Psychological Bulletin

HARRY HELSON, Editor
University of Texas

CONTENTS

Published Bimonthly by the
American Psychological Association

VOL. 57, NO. 5 SEPTEMBER 1960

The *Psychological Bulletin* contains evaluative reviews of research literature and articles on research methodology in psychology. This JOURNAL *does not publish reports of original research or original theoretical articles.*

Manuscripts should be sent to the Editor, Harry Helson, Department of Psychology, University of Texas, Austin 12, Texas.

Preparation of articles for publication. Authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (1957 Revision). Special attention should be given to the section on the preparation of the references (pp. 50–60), since this is a particular source of difficulty in long reviews of research literature. *All copy must be double spaced, including the references.* All manuscripts should be submitted in *duplicate*, one of which should be an original typed copy; author's name should appear only on title page. Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Reprints. Fifty free reprints are given to contributors of articles and notes.

ARTHUR C. HOFFMAN
Managing Editor

HELEN ORR
Promotion Manager

Communications—including subscriptions, orders of back issues, and changes of address—should be addressed to the American Psychological Association, 1333 Sixteenth Street N.W., Washington 6, D. C. Address changes must reach the Subscription Office by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $8.00 (Foreign $8.50). Single copies, $1.50.

PUBLISHED BIMONTHLY BY:

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Menasha, Wisconsin

and 1333 Sixteenth Street N.W., Washington 6, D.C.

Entered as second class matter at Washington, D.C., and Menasha, Wisconsin. Second-class postage paid at Washington, D.C., and at additional mailing offices. Printed in U.S.A.

VOL. 57, NO. 5 SEPTEMBER 1960

Psychological Bulletin

EXTRAVERSION-INTROVERSION AS A DIMENSION OF PERSONALITY: A REAPPRAISAL

PATRICIA M. CARRIGAN[1]
University of Michigan

Few personality constructs have remained as controversial and as productive of research over the years as extraversion-introversion. First established by Jung (1923) as a "dimension" of the normal personality, the construct has since been extended in many directions; it has been linked with physiological processes and morphology, with perceptual and cognitive behavior, with sociocultural phenomena, with physical and psychopathological disorders of one sort and another. Early attempts to demonstrate these relationships produced little in the way of definitive results; researchers began to doubt the validity of the construct, and in the early forties, it looked for a time as though extraversion-introversion had had its day. Like the proverbial bad penny, however, the construct has continued to turn up, notably in factor analytic studies, and over the past decade it has gradually been reinstated as an important focus in personality research. In a review of factorial studies of personality, Eysenck (1953) observed that although

> the picture is not as clear as one might wish . . . its main outlines are becoming more and more definite. . . . At the type level, i.e., at a level where concepts are based essentially on the intercorrelations between traits, three main dimensions appear to have been established: Neuroticism, Extraversion-Introversion, and Psychoticism. These three dimensions appear to be relatively orthogonal to each other, and also to "g" (Thurstone's second-order factor of cognitive functioning) (p. 318).

Eysenck's conclusions as they apply to extraversion-introversion embrace two issues of longstanding concern—briefly, the unidimensionality of the construct, and its relationship to "neuroticism" or, more broadly, adjustment. These issues were not adequately resolved at the time of Eysenck's review; they have gained importance in the years since, as a result of renewed interest in extraversion-introversion. In this paper, the two issues will be examined in the light of more recent evidence, in an attempt to clarify the current status of extraversion-introversion as a personality dimension.

THE ISSUES

Is extraversion-introversion a unitary dimension? Doubt concerning the unidimensionality of extraversion-introversion was a natural consequence of the conflicting results of early research; it was reinforced by the repeated finding of low to moderate correlations (averaging about

[1] The writer wishes to express her appreciation to E. Lowell Kelly for his valuable advice and assistance in the preparation of this paper, and to Warren T. Norman and Richard D. Mann for their reading of the manuscript.

.35) between various measures of the so-called dimension (Bernreuter, 1934; Guilford & Hunt, 1932; Hovey, 1929; Moore & Steele, 1934; Stagner, 1932; Vernon, 1938). While recognizing that the measures were partly at fault, investigators began to suspect, in addition, that they were not dealing with a single dimension.

Unidimensionality is clearly implied in Eysenck's conclusions, above; in support of his position, he points to factors in rating and questionnaire studies (and a few objective and projective test analyses) which, though bearing different names, seem to reflect extraversion-like characteristics. However, many inconsistencies can be found in the factors, and empirical evidence for their identity is virtually nonexistent.

In the past few years, many psychologists have become increasingly convinced that extraversion-introversion is an important dimension of personality; yet there is curiously little agreement as to its essential nature. Cattell (1957b), for example, has presented evidence to indicate that extraversion-introversion is largely of environmental origin; Eysenck (1956a) is as firmly convinced by his research that heredity plays a major role. The persistence of such discrepancies strongly suggests that extraversion-introversion may *not* be a single dimension after all, and that the impression of unity may have resulted from the too hasty attachment of a familiar label to a number of quite different dimensions, on the basis of superficial similarity.

Is extraversion-introversion independent of adjustment? The relationship between extraversion-introversion and adjustment was argued at length by early theorists, some following Jung (1923) in maintaining the independence of the two dimensions, others sharing Freud's (1920) belief that introversion was a forerunner of neurosis. Researchers generally accepted Jung's formulation, but ran into difficulty when it came to measuring the two dimensions. Guilford (1934) pointed up the problem, calling attention to

> the very troublesome situation found by those who construct tests of IE [introversion-extroversion] and of "neurotic tendency," a difficulty in keeping the two types of tests from correlating significantly with one another (p. 343).

Again, the measures were suspect, but with repeated attempts to improve them, measures of extraversion continued to correlate as highly with adjustment measures as they did with each other (Bernreuter, 1934; Vernon, 1938). Thus, the possibility of an intrinsic relationship between the two dimensions could not be ruled out.

The problem has been less apparent in factorial research, where, as Eysenck has noted, orthogonal factors resembling extraversion-introversion and neuroticism frequently appear in the same analyses. In many instances, however, the characteristics associated with "introversion" continue to have a strong maladjustive flavor.

Clarification of these issues must be sought in multivariate research, examined in the light of well-defined criteria for unidimensionality and factorial independence. The following criteria appear useful; they guide the presentation of evidence, below, and provide a framework for subsequent evaluation.

1. If extraversion-introversion is a major, unitary dimension of personality, (*a*) it should be represented as a factor in all measures and media covering the personality domain, and (*b*) the factors so obtained should be interrelated.

2. If extraversion-introversion and adjustment are independent dimen-

sions, (*a*) factors corresponding to the two dimensions should be uncorrelated, and (*b*) to the extent that the same variables appear on factors of extraversion-introversion and adjustment, indicators of "good" and "poor" adjustment should as frequently be associated with extraversion as with introversion.

The Evidence

In the present discussion, attention will be focused on research which has appeared since Eysenck's 1953 review. However, exceptions will be made in the case of earlier studies which are particularly relevant to the issues outlined above. The evidence to be considered comes chiefly from the researches listed (with code symbol and reference citation) in Table 1.

Analyses of Ratings

Few factorial rating studies have been published in recent years; the one contribution of interest here is a second-order analysis by Cattell (CaD) of "life-record" data, based on observations of behavior in life situations. Cattell's second-order extraversion factor has positive loadings on F, Surgency (.70); E, Dominance (.54); A, Cyclothymia (.38); and H, Parmia (.17), the latter associated with gregarious sociability and impulsiveness. At the introvert pole, the factor is defined by M, Autia (−.54), linked with extreme subjectivity and "inner mental life."

The results of this analysis are especially noteworthy, in view of Cattell's longtime insistence that extraversion-introversion (E-I) was nothing more than a broad cluster of related trait elements, and, as such, not a very useful construct (Cattell, 1945, 1946, 1950). Having discovered second-order E-I factors in rating and questionnaire data, he now suggests that

> it is perhaps worth while to make a determined attempt to rescue the label "extravert-vs-introvert" from the scientific disrepute and uselessness into which it has fallen through popular adoption (Cattell, 1957b, p. 267).

Analyses of Questionnaires

Much of the factorial research relevant to E-I is based on questionnaires which evolved from a series of early factorial studies by Guilford and Guilford (1934, 1936, 1939a, 1939b). Among these questionnaires are Guilford's Inventory of Factors STDCR (1940); the Guilford-Martin Inventory of Factors GAMIN (1943a); the Guilford-Martin Personnel Inventory (1943b); and the Guilford-Zimmerman Temperament Survey, or GZTS (1949), which incorporates 10 factors from the preceding inventories

Analyses of the Guilford Questionnaires

It can be seen in Table 2 that analyses of the Guilford questionnaires have consistently yielded E-I factors defined by a similar pattern of variables. Factors obtained by Denton and Taylor (De) and by North (No), in analyses of the STDCR inventory, have their principal loadings on R, Rhathymia (freedom from care) and S-, Social Extraversion.[3] R and S- also appear on Lovell's factor (Lo) along with G, General Activity, and A, Ascendance, from the GAMIN inventory. The same four variables are distributed on three of the factors obtained by Thurstone (Thu) in a reanalysis of Lovell's data. A second-order analysis of Thurstone's matrix by Baehr (Ba) pulls together R, S-, G, and A on an extraversion-like factor, Primary Function, which is defined by Thurstone's Factors VII, Impulsivity (.85) and V, Dom-

[3] Denton and Taylor's factor also has a loading of .29 on an objective test factor called Verbal Versatility.

TABLE 1

MAJOR CHARACTERISTICS OF MULTIVARIATE STUDIES RELEVANT TO EXTRAVERSION-INTROVERSION (E-I)

Investigator	Symbol	Sample	Number and Type of Variables[a]	Type of Analysis[b]	Number of Factors
Bachr (1952)	Ba	(Same as Thu, below)	9 Q	Second-order, oblique	6
Becker (1959)	Be	32 male, 30 female college students	25 T[c] 15 Q	First-order, orthogonal	8
Cattell (1955)	CaA	250 USAF pilot trainees	64 T	First-order, oblique	15
	CaB	500 USAF pilot trainees	113 T 15 Q	First-order, oblique	16
Cattell (1956b)	CaC	181 male & female college students, 227 USAF trainees	15 Q	Second-order, oblique	4
Cattell (1957b)	CaD	544 male & female college students	12 R	Second-order, oblique	6
Cook & Wherry (1950)	Co	111 naval submarine candidates	11 Q 6 T 2 C 1 R	First-order, orthogonal	6
Denton & Taylor (1955)	De	170 high school seniors	8 T 5 Q	Second-order, oblique	6
Eysenck (1956a)	Ey	104 twins (13 pairs each male identical & fraternal, female identical & fraternal)	34 T 11 PT 2 R 2 SR 1 Q	First-order (rotational criteria not specified)	6
Foster (1955)	FoA	54 state highway patrol officers	8 PT 2 T 1 C	First-order, orthogonal	4
	FoB	28 college student volunteers	8 Q 4 PT 4 I 3 T	First-order, orthogonal	5
Franks, Soueif, & Maxwell[d]	Fr	100 male, 100 female adult volunteers	7 Q	First-order, orthogonal[e]	3
Guilford & Zimmerman (1956)	Gu	(Same as Lo, below)	69 Q	First-order, orthogonal (r_{tt}, cosine-pi approximation)	18
Heron (1954)	He	80 male unskilled factory workers	19 T 4 I 1 C 1 R	First-order, orthogonal (Burt's simple summation)	4
Hildebrand (1958)	Hil	95 male neurotics	15 T 6 Q 1 I	First-order[e]	4
Himmelweit, Desai, & Petrie (1946)	Him	64 male surgical patients	16 T 1 Q 1 C	First-order orthogonal (r_{tt}; unrotated)	7
Karson & Pool (1957b)	KaA	71 maladjusted USAF officers	80 Q	Correlational	—
Karson & Pool (1958)	KaB	71 maladjusted USAF officers	16 Q	Second-order, orthogonal	6
Kassebaum, Couch, & Slater (1959)	Kas	160 male college freshmen	32 Q	First-order, orthogonal	3
Lovell (1945)	Lo	122 male, 78 female college students	13 Q	Second-order, orthogonal	6

[a] Classified as follows: C, clinical observation; I, interest or attitude inventory; PT, projective test; Q, questionnaire; R, behavioral rating; SR, self-rating; T, objective test.
[b] Unless otherwise noted, all factorizations began with centroid analyses based on Pearson r, factors subsequently rotated for simple structure.
[c] Plus 7 retest measures, 5 random variables, 6 items of background information.
[d] Unpublished study, 1958.
[e] Not rotated for simple structure; rotational criteria described in text.

TABLE 1 (Continued)

Investigator	Symbol	Sample	Number and Type of Variables	Type of Analysis	Number of Factors
R. D. Mann (1958)	MaA	100 male college students	26 Q 8 T 5 I 1 R	Second-order,[f] orthogonal	7
R. D. Mann[g]	MaB	100 male college students	26 Q	Second-order, orthogonal	5
	MaC	100 female college students	26 Q	Second-order, orthogonal	5
Nelson & Shea (1956)	Ne	19 male, 33 female college students	15 Q	Correlational	—
North (1949)	No	135 male, 15 female college students	5 Q	Second-order, orthogonal	2
Royal (1950)	Ro	100 male college students	12 PT 3 Q	Correlational (r_{pb})	—
Scheier & Cattell (1958)	Sc	86 male college students	90 T 17 Q 6 R	First-order, oblique	15
Singer, Wilensky, & McCraven (1956)	Si	100 male schizophrenics	9 PT 9 T 4 R 1 SR	First-order, oblique ($r\phi$); second-order, oblique	4 2
Thornton & Guilford (1936)	Tho	75 male, 25 female college students	5 Q 5 PT	Correlational	—
Thurstone (1951)	Thu	(Same as Lo, above)	13 Q	First-order, oblique	9
Tyler (1951)	Ty	107 female graduate students	15 Q	First-order, orthogonal & oblique	5
Welsh (1956)	We	150 male VA medical & surgical patients	16 Q	First-order, orthogonal (unrotated)	3
Wheeler, Little, & Lehner (1951)	Wh	112 male college students (matrix 1)	12 Q	First-order, orthogonal	4
Williams & Lawrence (1954)	Wi	100 male VA neuropsychiatric patients	17 PT 14 Q 1 T	First-order, orthogonal & oblique (r_t for T & PT variables)	4
Wood (1957)	Wo	56 male & female college students	18 Q	Correlational	—

[f] With respect to Q variables only. However, factors discussed here have no important loadings except on Q variables, hence are essentially second-order factors.
[g] Unpublished analyses, Univer. of Michigan, 1959. Based on data obtained by Weitzenhoffer (1956).

inance (.80). However, Thurstone's first factor, Reflectiveness, with its principal loading on T, appears instead on Baehr's Emotionally Unstable factor. From this analysis—and from the preceding ones—it looks as if T, Thinking Introversion, is essentially a maladjustment factor,[9] and that the core of E-I as measured by the Guilford questionnaires consists of Factors R, S, G, and A.

A question about the relationship of R to extraversion has been raised by Guilford and Zimmerman, who have recently carried out another analysis (Gu) of Lovell's data. In order to have several variables representing each factor, they divided each of the factor scales into three or more short "tests," by sorting the items into apparently homogeneous subgroups. Sixty-nine "tests" or variables were obtained in this manner; another—the subject's sex—was added. The matrix of intercorrelations for the 70 variables yielded fairly good approximations of the 13 original questionnaire factors, along with a second C factor—C2—and four residuals. Minor changes in meaning

[9] However, its GZTS counterpart, Thoughtfulness, loads several extraversion-like factors obtained in joint analyses of the Guilford and Cattell questionnaires, discussed subsequently.

TABLE 2

QUESTIONNAIRE FACTORS: GUILFORD AND CATTELL FACTORIAL QUESTIONNAIRES

Variable[a]	Factor Identification (Analysis & Factor in Series)															
	De V	Lo I	No II	Thu I-	Thu V	Thu VII	CaC II	KaB II	Be V[c]	MaA III	MaA IV	MaB III	MaB IV-	MaC II	MaC III-	MaC IV
Guilford factors[b]																
G, General Activity		73		−02	−04	60				60	04	62	02	64	−07	29
R, Restraint										−07	−64	−25	−67	−48	−51	−26
(R, Rhythmia)	71	71	75	41	07	45			81							
A, Ascendance		70		−03	55	00				56	02	83	−21	77	−19	−15
S, Sociability										70	06	77	00	80	11	−22
(S, Social Introversion)	−51M	−70M	−75M	−29	−42	−06M										
E, Emotional Stability										18	10A	19	00A	29	03	01A
(C, Cycloid Tendency)	42M	−02M	01M	−26	00	05M										
(D, Depression)	01M	−23M	−24M	−35	−12	01M										
O, Objectivity		25A		08	00	13				06	04A	09	05A	19	−17	−05A
F, Friendliness										07	−31A	−28	01A	−25	07	04A
(Ag, Agreeableness)		−08A		−03	−01	−03			−17A							
T, Thoughtfulness										04	−51	23	−72	−08	−58	−17
(T, Thinking Introversion)	−03M	−08M	−03M	−76	00	−02M			−61							
P, Personal Relations										11	−11A	−01	05A	03	−08	−10A
(Co, Cooperativeness)		06		03	07	−03			09A							
M, Masculinity		−02		00	01	−04			−07	06	08A	09	13A	19	−41	−04A
−, (I, Inferiority)		−38M		05	−04	−17										
−, (N, Nervousness)		00M		−01	00	−05M										
Cattell factors																
A, Cyclothymia							45	51		22	05	31	08	30	23	−04
B, General Intelligence								03	−08	03	−13	−15	05	−02	−15	02
C, Emotional Stability							10A	24A		11	−02A	−11	23A	−03	−04	08A
E, Dominance							−01	54		36	28	51	06	55	−30	23
F, Surgency							48	75		31	57	49	50	55	27	09
G, Super-Ego Strength							−01	−06	−39	22	−28	20	−45	−11	06	−45
H, Parmia							33A	84A		68	09	85	00	86	01	−12
I, Premsia							01	−23M	−18	10	−35	−06	−37	−11	03	25
L, Protension							−13M	−02M		−22	33M	11	−05M	−10	−10	01M
M, Autia							−51	02M	−15	−08	13M	11	−11M	30	−10	33M
N, Shrewdness							−09	11	06M	21	34M	05	−05	02	−18	−01
O, Guilt-Proneness							−07M	−34M		−30	−17M	−29	09M	−15	−04	19M
Q1, Radicalism							−42	02M	−08	09	02	01	−05	09	−52	17
Q2, Self-Sufficiency							−38	−23		−30	11	−32	−23	−34	−38	31
Q3, Will Control							−01A	−19A		06	−34	−15	−51A	01	−07	−59A
Q4, Ergic Tension							−09M	−06M		05	01M	−03	07M	−07	−03	−14M

Note.—In this and subsequent tables, decimal points omitted, factor loadings reported to two decimal places. Vacant cells indicate variables not represented in study. A and M (attached to factor loadings) indicate variables appearing on "adjustment" factors ($\geq .30$) in same analysis, and direction of loading (A = adjustment, M = maladjustment).

[a] Loadings reflected where necessary to agree with direction indicated by factor scale title.

[b] Listed by title in Guilford-Zimmerman Temperament Survey. Earlier equivalents, if different, given in parentheses. Note change in direction for R and S.

[c] Additional loadings: Maudsley Personality Inventory E, Extraversion, 82; N, Neuroticism −28M; 16 PF second-order Anxiety −14M; 16 PF second-order Extraversion (sum A, E F, and H), 68.

were indicated for several factors, and rather substantial ones for R. Half of the "tests" from the R scale went to other factors: reticence to A, impulsivity to C2, and rapport with the environment to O. In view of these modifications, particularly the last, Guilford and Zimmerman have questioned the relationship of R to Jung's extraversion, with which it has generally been identified. However, the remaining attributes of R—carefreeness, unconcern, and liking for action, along with the cheerfulness and energy formerly associated with Factors D and G—seem in a broad sense, at least, to be consistent with extraversion.

Analyses of the 16 PF Test

Cattell's E-I questionnaire factor emerged from a second-order analysis (CaC) of the Sixteen Personality Factor Questionnaire, or 16 PF test (Cattell, 1957a). This factor, shown in Table 2, is similar to the previously discussed rating factor, differing chiefly in the omission of E, Dominance, and the addition at the introvert pole of two primary factors unique to questionnaire data—Q1, Radicalism, and Q2, Self-Sufficiency.

The 16 PF extraversion factor obtained by Karson and Pool (KaB) resembles Cattell's in F, Surgency, and A, Cyclothymia, but the two factors are otherwise quite different. As seen in Table 2, Karson and Pool's factor adds E, drops Q1, and has a negligible Q2 loading; more important discrepancies are found in H, Parmia, and M, Autia. Cattell's factor has its highest loading on M, and a relatively small one on H; Karson and Pool's E-I factor, on the other hand, has its highest loading on H and no loading on M, which appears on their anxiety factor (.72). Similarly, M contributes little to the extraversion-like factors obtained in joint analyses of the Guilford and Cattell questionnaires, but in three of these analyses (MaA, MaB, MaC) it has substantial loadings—.44, .59, .54—on maladjustment. Furthermore, Wood's 16 PF intercorrelations (Wo) show M to be virtually uncorrelated with the other extraversion primaries, but closely related to the major components of Cattell's second-order anxiety factor (L, O, Q3, Q4). Contrary to Cattell's results, then, these studies suggest that M is primarily a maladjustment factor.

The various studies do differ in several respects, and while the discrepant results are not adequately accounted for by these differences, it is well to mention them. In the first place, Cattell's factors are oblique, Karson and Pool's factors—and factors from the joint analyses—are orthogonal. The use of different rotational criteria might be expected to result in somewhat different factor patterns; it is not a sufficient explanation, however, for the correlation matrices themselves are quite dissimilar. Cattell's matrix shows M, for example, to be a relatively independent factor, having its highest correlation (−.36) with F. On the other hand, Karson and Pool, Mann, and Weitzenhoffer (1956)—whose matrices were used in Mann's B and C analyses—found M to be substantially correlated with several factors, notably the adjustment primaries. It should be noted, too, that Cattell's matrix consists of correlations between the primary *factors*, the others of correlations between *factor scores*. However, for the sample on which Cattell's analysis is based, the correlations between factor scores do not differ greatly from the primary factor intercorrelations (Cattell, 1957a). Finally, the various analyses are based on somewhat different pop-

ulations—i.e., Karson and Pool's on Air Force personnel, Mann's on college students, Cattell's on a combined group of college students and Air Force trainees. On the basis of population differences, then, it would be expected that the greatest discrepancies would be found between Karson and Pool's analyses and Mann's. Quite the contrary, these studies yielded the most comparable intercorrelations, and the Karson and Pool E-I factor is closely paralleled by one factor from each of Mann's analyses (MaA and MaB III, MaC II). Moreover, these studies unanimously fail to support M as a major E-I variable. Thus, while some of Cattell's primary factors—notably F and H—seem well-established as nuclear parts of the extraversion pattern, the role of M remains unclear.

Joint Analyses: Guilford and Cattell Questionnaires

Inasmuch as the questionnaires of Guilford and Cattell cover a wide range of personality characteristics, it might be expected that the measures would overlap to some extent, and that the two sets of extraversion factors would be closely related. The nature of the relationship can be seen in Table 3, in the GZTS-16 PF intercorrelations obtained by Weitzenhoffer. It is interesting to note that the questionnaire scales with consistently high loadings on E-I factors—Guilford's R, S, G, and A, and Cattell's F and H—form a highly correlated "cluster," and that except for E, Dominance, the remaining extraversion primaries from the two inventories are only tangentially linked with the cluster.

Of greater interest is what happens to the cluster when the intercorrelations for the Guilford and Cattell scales are jointly factored. Relevant factors from Mann's analyses (MaA, MaB, MaC), and Becker's (Be), appear in Table 2. One of the first things to be noted is that only one of the joint analyses yielded a factor which clearly corresponds to the cluster described above: Factor II in Analysis MaC, which has its principal loadings on 16 PF E, F, and H, and GZTS G, R-, A, and S. The MaA and MaB analyses split the cluster and distributed its variables on two factors—Factor III, Social Extroversion, which combines GZTS S, G, and A with 16 PF H and E, and Factor IV, Lack of Self-Control, which links GZTS R- (and T-, one of the "fringe" variables) with 16 PF F. Becker's factor seems most closely related to the latter, by virtue of its loadings on Guilford's R- and T- and 16 PF G-, Lack of Internal Standards. Unfortunately, comparison is hindered by the fact that 16 PF A, E, F, and H are represented by a single score in Becker's analysis (see footnote, Table 2). Finally, looking again at the MaC analysis, it will be noted that Factor II, despite its sizeable loadings on all of the cluster variables, is most heavily weighted by GZTS S, G, and A, and 16 PF H; in short, it is most similar to the Social Extroversion factors from the MaA and MaB analyses. MaC Factor III-, with its GZTS R- and T- loadings, and Factor IV, defined principally by 16 PF G- and Q3-, Lack of Will Control, may be a further split of the Lack of Self-Control factors obtained in the MaA and MaB analyses.

From Mann's analyses, then, it appears that two or more factors are required to account for the intercorrelations between E-I variables from the Guilford and Cattell questionnaires. Moreover, the factors show remarkably little overlap; only F, Surgency, has loadings as great as .30 on the two factors from the MaA and MaB analyses. It would seem, there-

TABLE 3

INTERCORRELATIONS BETWEEN E-I VARIABLES: GZTS AND 16 P. F. QUESTIONNAIRES
(From Weitzenhoffer, 1956)

Variable	GZTS				
	G	R	A	S	T
16 P. F.					
A	16	−15	19	28**	−05
	07	−27**	20*	29**	−05
E	29**	−26**	44**	37**	*17*
	39**	−07	47**	34**	−04
F	30**	−60**	26**	44**	−17
	40**	−54**	28**	41**	−29**
H	47**	−36**	74**	74**	*06*
	45**	−40**	69**	81**	−02
M	*10*	06	*04*	−18	32**
	*33***	*−30***	*17*	*05*	15
Q1	*13*	29**	*19*	−14	37**
	06	06	*11*	*06*	40**
Q2	−19	23*	−11	−45**	28**
	−08	26**	−28**	−44**	17

Note.—Intercorrelations based on 100 males, 100 females, respectively. Italicized coefficients indicate reversals of expected sign.
* Significant at .05.
** Significant at .01.

fore, that these factors represent relatively distinct dimensions.

As to the nature of the dimensions, Mann (1958), in a discussion of his MaA factors, has suggested the possibility that

> Factor III corresponds to the American conception of extroversion, with its emphasis on sociability and ease in interpersonal relations, while Factor IV corresponds to the European conception of extroversion, with its emphasis on impulsiveness and weak superego controls (p. 108).

Mann's distinction appears to be a valuable one; perhaps, however, it can be more precisely tied down in terms of the major variables defining the two factors.

Looking first at Social Extroversion, the vitality and enthusiasm associated with GZTS G, the aspiration to leadership and interpersonal interaction reflected in A and 16 PF E, the seeking of (and pleasure in) social contacts described by GZTS S and 16 PF H, all appear to be ingredients of response to the environment and its "objects," i.e., people. A dimension described by these variables might then be broadly conceptualized as one of response to external stimuli, with the extremes characterized as approach vs. avoidance. Thus defined, Social Extroversion would seem to approximate Jung's (1923) conception of extraversion, the essence of which is the relative importance accorded the "object" and objective events. The negative pole of the factor might likewise be identified with Jung's introversion—emphasis on the self and inner, subjective processes—to the extent that avoidance of the external world can be viewed as a consequence of such self-preoccupation.

Mann's Lack of Self-Control factor, on the other hand, suggests a very different conception of E-I. Among the variables defining this factor, GZTS R- contrasts happy-go-lucky unconcern with seriousness and

self-control; T- is associated with mental disconcertedness, as opposed to reflectiveness and self-observation; 16 PF F reflects carefreeness vs. introspectiveness and brooding;[4] G- is associated with lack of dependability and indolence, as opposed to perseverance and conscientiousness; Q3- contrasts laxity with control. The essence of these characteristics seems to be their relevance to the handling of impulses; the dimension they describe might be thought of as one of response to stimuli arising from within. Viewed in this way, Mann's Lack of Self-Control factor is readily identified with Eysenck's conception of E-I. In his *Dynamics of Anxiety and Hysteria*, Eysenck (1957) characterizes the neurotic extravert as undersocialized (schematically, id +ego>super-ego); the neurotic introvert, on the other hand, is described as oversocialized (super-ego +ego>id). An empirical link with Eysenck's viewpoint is provided by the high R loading on Lack of Self-Control. Eysenck considers R to be a good measure of his dimension; he has used it both as a research criterion and as the basis for the Extraversion (E) scale in the Maudsley Personality Inventory (Eysenck, 1956b). Of further interest is the fact that until recently, at least, Eysenck has been unwilling to include sociability as part of his extraversion constellation. In view of the independence of Social Extroversion and Lack of Self-Control, it appears that he may have been quite correct.

Several implications can be drawn from the joint analyses reported here. One concerns the relationship between E-I and adjustment. On the basis of the MaA analysis, a good case can be made for identifying Social Extroversion as a factor of "well-adjusted" extraversion. It can be seen in Table 2 that the factor (III) tends to have positive loadings on variables associated with "good" adjustment, negative ones on variables related to maladjustment. On the other hand, the MaB and MaC counterparts (Factors III and II, respectively) do not reflect this tendency. It should be pointed out, however, that the highest-loading variables on Social Extroversion—GZTS S and 16 PF H—have small but consistently positive loadings on adjustment factors, in every analysis which included them. It cannot be denied, moreover, that in a culture such as ours, which places a high premium on interpersonal interaction, the characteristic avoidance of such interaction—associated here with social *intro*version—might be considered maladaptive.

On the other hand, if Mann's Lack of Self-Control factor is correctly identified with Eysenck's dimension, it would appear that *both* extremes of this factor are linked with maladjustment. Presumably the individual whose ego mediates a more harmonious relationship between the expression and control of impulses—i.e., the individual falling near the middle of the dimension—would be better adjusted than individuals at either extreme. However, to the extent that society rewards self-control and conformity to cultural standards, the factor might be looked upon as contrasting maladjusted extraversion with well-adjusted introversion. The latter interpretation is favored by the MaA analysis, where Factor IV tends to have positive loadings on "maladjustment" variables, negative ones on variables reflecting "good" adjustment (see Table 2). Again, however,

[4] F also contrasts enthusiasm, cheerfulness, and talkativeness with incommunicativeness—a contrast more relevant to response to the environment. That F loads both factors is thus not surprising.

the MaB and MaC analyses do not concur.

Further implications stem from the independence of the two factors. While the relationship to adjustment requires further clarification, if it should turn out that Social Extroversion and Lack of Self-Control *do* reflect well-adjusted and maladjusted extraversion, respectively, the lack of overlap in the two factors might suggest that extraversion and introversion are differentially manifested in individuals falling at opposite ends of the adjustment continuum. Discrepancies between the MaA and MaB factors, based on male subjects, and the MaC factors, based on female subjects, suggest further a qualitative sex difference in E-I. A final implication concerns the unidimensionality of E-I. In view of the independence of Mann's factors, it is quite clear that the dimensions they represent cannot be subsumed under the same label.

Analyses of the MMPI

With a few exceptions (Abrams, 1949; Cottle, 1950; Wheeler, Little, & Lehner, 1951 [Matrix 2]), factorial studies of the Minnesota Multiphasic Personality Inventory (MMPI) have consistently yielded bipolar factors with contrasting loadings on *Ma*, Hypomania, and *D*, Depression. That these factors (Table 4) may be related to E-I is suggested by several analyses, in which the MMPI clinical scales have been supplemented by various "personality" scales developed for the inventory.

Two factors from Tyler's analysis (Ty) are relevant. As Table 4 shows, Factor II, a "hysteroid" conflict factor, adds to the *Ma-D* contrast a dimension of responsibility (*Re*) not uncommonly associated with introversion. The appearance of *Hy*, Hysteria, at the "introvert" extreme requires comment, however. According to Eysenck's theory,[5] hysteria is associated with extraversion, and the negative *Hy* loading—here, and on several other factors in Table 4—would thus seem to be inconsistent. It will be seen later, however, that the MMPI *Hy* scale is essentially unrelated to Eysenck's E-I dimension. Of the remaining variables defining Tyler's second factor, the prominence of *Pt*, Psychasthenia, and *Sc*, Schizophrenia, might suggest that the factor is one of "maladjusted" introversion, but the overall resemblance of the factor to E-I is not impressive.

Tyler's third factor, Social Aggressiveness, differs somewhat in the orthogonal and oblique rotations. The oblique factor links *Ma* with *Do*, Dominance, and *St*, Social Status, and has its highest loading on another scale suggestive of Eysenck's extraversion—*Pd*, Psychopathic Deviate. *D* does not appear on the factor, but the variables defining the negative pole do not seem inconsistent with introversion. The orthogonal factor, on the other hand, is less well defined, and the substantial *Pt* and *Sc* loadings suggest that it would have to be looked upon as a factor of "maladjusted" extraversion. In general, Tyler's analysis seems to confirm the presence of an extraversion-like dimension in the MMPI, but the exact nature of the dimension is by no means clear.

Welsh's analysis (We) is not strictly comparable to other MMPI analyses. It is based chiefly on *prime scales*—modified versions of the original scales containing no multiple scored items and, hence, not subject to the spurious intercorrelation introduced by item overlap. Several other special scales are included:

[5] The reader unfamiliar with Eysenck's theory will find a brief discussion later in this paper.

TABLE 4

QUESTIONNAIRE FACTORS: MMPI

Variable	Factor Identification									Factor Identification	
	Co III	Fr II[a]	Fr III[b]	Kas II-	Ty II-[c]	Ty III[c]	We II-	Wh IV-		Kas II- (con'd.)	We II- (con'd.)
L, Lie Scale	−34			−40				—	*A*, Anxiety	−02M	−11M
F, Validity Scale	02M			−02M				—	*Ac*, Achievement via Conformance	−28A	
K, Suppressor Scale		46	06A	−07A			−13A	—	*As*, Achievement via Independence	−17A	
Hs, Hypochondriasis	−23M			01M	−29, —	33, 32M	−18M	—	*Fm*, Feminine Masochism	−37M	
D, Depression	−48			−40M	−35, −31	—M	−39A	−60M	*Gm*, Maladjustment		−49M
Hy, Hysteria	−43	79	30	03	−39, −34	—, 39	−46A	—	*Ie*, Intellectual Efficiency	12A	
Pd, Psychopathic Deviate	−38M	33	30M	20M	—	60, 58	−13M	—	*Im* Impulsivity	56M	
Mf, Masculinity-Femininity	01M			−10M	—	—	−38	—	*Ja*, Judged Anxiety		−23M
Pa, Paranoia	−17			−06M	−41, −43	27, 56A	−34	—	*Lp*, Leadership	26A	
Pt, Psychasthenia	10M	15	03M	00M	−44, −45	55, —M	−05M	—	*M*, Mania		45M
Sc, Schizophrenia	14M			11M	−30, —	50, 36M	−02M	—	*Of*, Originality	−28A	
Ma, Hypomania	56M			63M	32, 31	54, 44	52M	60	*R*, Repression	−69	−88
Do, Dominance				20A	—	—, 34A			*Rp*, Role-playing Ability	15A	
Pr, Prejudice				14M	32, 30	—, −28M			*Sp*, Social Presence	44A	
Re, Social Responsibility				−46A	−49, −48	−48, —			*Sy*, Sociability[d]	55A	
Si, Social Introversion				−58M	—	—, −33M	−24M		*To*, Tolerance	−19A	
St, Social Status				42A	—	—, 33A					

Note.—In this and subsequent tables, dash (—) indicates variables for which no loading is reported, though included in study.

[a] Additional loadings: STDCR *C*, −04; *D*, 01; *R*, −01.

[b] Additional loadings: STDCR *R*, 76; *C*, 14M; *D*, −12M.

[c] Orthogonal and oblique loadings, respectively.

[d] Formerly Social Participation.

Gm, consisting of items scored for at least three MMPI scales and thought, therefore, to reflect some general MMPI dimension, presumably maladjustment; *Ja*, a rational scale of anxiety; and three empirical scales—*A*, Anxiety, *M*, Mania, and *R*, Repression—made up of items found to differentiate between subjects scoring at the high and low extremes on the *Gm*, *Ma'*, and *D'* scales, respectively.

Welsh's second factor contrasts *Ma'* and *M* with *D'* and *R*, and has a small loading on *Si'*, Social Introversion; it also brings in at the "introvert" pole the rationally-derived maladjustment scales (*Gm*, *Ja*), although *Pt'* and *Sc'* do not appear on the factor. Inasmuch as Welsh's factors are unrotated, the analysis as it stands is not very satisfactory. It does, however, provide added evidence for an E-I dimension in the MMPI—a dimension which apparently transcends item overlap in the scales.

By far the most impressive results are those from Kassebaum, Couch, and Slater's analysis (Kas). Their second factor, Introversion-Extraversion, links *D* with *R*, *Si*, *Re*, and *L*, Lie Score; *Ma* appears at the extravert pole, along with *Im*, Impulsivity, and several scales suggestive of a "social" orientation. The factor thus incorporates several scales associated with the other factors in Table 4, and adds a number of "personality" scales which further identify it as an E-I factor.

In their discussion of the factor, Kassebaum et al. have called attention to the fact that two of the scales defining introversion—*R* and *L*—consist solely of items scored for a "False" response; a third "introversion" scale, *D*, likewise has a preponderance of "False" items, whereas *Ma*, associated with extraversion, contains significantly more "True" items. On the basis of these facts, the writers suggest the possibility that

> what we have labeled extraversion is associated with a general tendency to agree with any item whatever the content, while what we have called introversion involves a converse tendency to disagree or mark False (Kassebaum, Couch & Slater, 1959, p. 230).

The extent to which such a "response set" may be involved in the various E-I questionnaire factors will be considered later.

It can be seen in Table 4 that most of the MMPI scales contribute substantially to maladjustment factors. Hence, in Kassebaum, Couch, and Slater's analysis, all of the E-I variables except *R* and *L* have loadings of .48 or above on Factor I, Ego Weakness. The nature of the relationship between these two "dimensions" is clarified to some extent by a further step in the analysis. Kassebaum and his colleagues reasoned that if their first two factors were correctly interpreted as Ego Weakness (maladjustment) and Introversion-Extraversion, it should be possible to identify more precisely the characteristics of "normal" and "disturbed" extraversion and introversion by rotating the axes 45 degrees and redefining the factors in their new positions. The axes were shifted accordingly, yielding two *fusion factors*, so named because they were thought of as combinations of the primary reference axes. Fusion Factor A, contrasting maladjusted introversion with normal extraversion, was labeled Social Withdrawal vs. Social Participation. It had its principal loadings on *Si*, *D*, and *Fm*, Feminine Masochism, and, at the negative pole, on the "social" scales *Sp*, *Sy*, and *St*. Fusion Factor B, Impulsivity vs. Intellectual Control, contrasted maladjusted extraversion with well-adjusted introversion. Its largest loadings were on *Im*, *Ma*, and, negatively, on *Re*, *To*,

Tolerance, and *Ac*, Achievement via Conformance. As would be expected, the two factors shared substantial loadings on a number of scales related primarily to maladjustment.

The conceptual similarity of the fusion factors to Mann's Social Extroversion and Lack of Self-Control is apparent. And, just as Mann's analyses split the cluster of extraversion variables from the GZTS and 16 PF questionnaires, the two fusion factors obtained by Kassebaum et al. show a clear separation of the MMPI scales contributing to their E-I factor, with the exception of *R*, Repression, which has sizeable loadings on both A and B. Again, there seems to be a clear implication that differences in adjustment may be associated with qualitatively different manifestations of extraversion and introversion.

Joint Analyses: MMPI and the Factorial Questionnaires

Relationships between the MMPI and the Guilford and Cattell questionnaires have been explored in several studies. A recent analysis by Franks, Souief, and Maxwell (Fr) is based on seven scales from the MMPI and STDCR inventories, selected as likely measures of E-I or neuroticism: Guilford's D, C, and R scales, and MMPI *K*, *Hy*, *Pd*, and *Pt*. The second and third factors obtained in the analysis are shown in Table 4; they were rotated orthogonally to Factor I, Neuroticism, so as to obtain a zero loading for R on Factor II. Thus rotated, Factor II has a large loading on *Hy*, smaller ones on *K* and *Pd*, and is identified as Extraversion-Introversion. Factor III, Rhathymia, has its principal loading on R.

This study raises several questions which, unfortunately, go unanswered.[6] Nowhere do the writers offer an explanation for the seemingly arbitrary rotation of Factors II and III. The deliberate elimination of R from the E-I factor is puzzling; Franks himself has used R as a measure of extraversion (Franks, 1956; Franks & Laverty, 1955; Laverty & Franks, 1956), and, moreover, the R scale was included in the present study because of its previously demonstrated relationship to the E-I dimension. The rationale for the interpretation of Factor II is equally unclear. The authors cite studies by Eriksen (1954a, 1954b) and by Eriksen and Davids (1955), which showed that college students obtaining high *Hy* and *Pt* scores, respectively, had certain characteristics in common with Eysenck's extraverted and introverted neurotics. However, these findings do not seem especially relevant; *Pt* has a small loading on Factor II, but in the same direction as *Hy*! Apparently, then, the interpretation of Factor II rests on Eysenck's association of hysteria and psychopathy with extraversion. Yet data from one of Eysenck's studies (1952) show that *Hy*—the highest-loading variable on Factor II—does not differentiate between hysterics, psychopaths, and anxiety states, the latter an "introvert" group; moreover, *Hy* correlated negatively (−.115) with the hysteria-anxiety dichotomy.

Franks et al. are undoubtedly wise to make a conservative interpretation of Factor III, in terms of its major variable, R. Nevertheless, the identification of this factor (rather than Factor II) with E-I would seem

[6] The writer was privileged to read and comment on Franks' manuscript some time ago, and understands that it is undergoing revision before being submitted for publication. Perhaps the issues raised here will be dealt with in the revision.

to be more defensible on the basis of existing evidence.

A broader picture of the relationships between E-I variables from the MMPI and the factorial questionnaires comes from two correlational studies—one by Nelson and Shea (Ne), using MMPI and the STDCR inventories, the other Karson and Pool's study (KaA) of the MMPI and 16 PF. Relevant coefficients from the two studies are reproduced in Table 5. It can be seen that only the *Si* scale is consistently related to the extraversion primaries from the two factorial questionnaires. *Ma* and *D* tend to correlate with the principal scales from the previously described extraversion "cluster" (Guilford's R and S, Cattell's F and H) but with no others; and the remaining MMPI scales—*K*, *Hy*, *Pd*, and *L*—have little in common with the factorial measures.

Karson and Pool's data shed further light on the nature of 16 PF Factor M, Autia. M correlates not only with *K*- and *Si*, but also with MMPI *Pt* (.48), *Sc* (.48), *Mf* (.47) and *F* (.46)—all scales which are linked with maladjustment (see Table 4). Thus, earlier indications that M may be essentially a maladjustment factor seem to be borne out here.

It is regrettable that the authors of these studies did not carry out factor analyses of their data. In Karson and Pool's publications, complex relationships between the scales are not readily disentangled by inspection of the matrix, although some clarification is provided in a separate article (Karson, 1958). Nevertheless, the two studies are of interest in providing an empirical link between MMPI *Ma*, *D*, and *Si* and the principal extraversion primaries from the factorial questionnaires.

Questionnaire Factors and Acquiescence

That E-I questionnaire factors may reflect certain response tendencies—as has been suggested by Kassebaum, Couch, and Slater in connec-

TABLE 5

INTERCORRELATIONS BETWEEN E-I VARIABLES: MMPI, STDCR, AND 16 PF QUESTIONNAIRES
(From analyses Ne and KaA)

Variable	MMPI[a]					
	K	*D*	*Hy*	*Pd*	*Ma*	*Si*
STDCR						
S		28*	−24	−15	−46**	69**
R		−39**	*−07*	*−06*	50**	−50**
16 PF						
A	13	−11	20	*−09*	00	−33**
E	23	−11	05	02	10	−27*
F	08	−26*	05	01	24*	−48**
H	27*	−24*	05	00	19	−69**
M	−48**	17	*16*	*29**	*17*	32**
Q1	−10	22	*30**	*19*	*11*	11
Q2	−22	01	−03	−02	*03*	32**

Note.—Italicize coefficients indicate reversals of expected sign.
[a] *L* not included in analysis Ne, did not correlate significantly with any 16 PF extraversion primaries.
* Significant at .05.
** Significant at .01.

tion with their MMPI factor—is a possibility which merits careful consideration. Evidence has been presented to show that an E-I dimension *can* be demonstrated in the various questionnaires. However, that evidence rests on the assumption that the questionnaire factors can be validly interpreted in terms of the "psychological meaning" of the variables which define them. If the covariation among the factor variables can be accounted for on some basis other than common meaning, the label extraversion-introversion would seem to be prematurely applied, and perhaps inappropriate.

In order to examine the susceptibility of the various "extraversion" scales to agreement response set, or acquiescence, the principal questionnaire variables have been listed in Table 6, along with the percentages of "agreement" and "disagreement" items they contain. Looking first at the Guilford scales, it can be seen that, in general, extraversion *does* tend to be associated with a higher percentage of agreement items, the only exceptions being GZTS S, which contains equal numbers of "True" and "False" items, and T, where the trend reverses.

An attempt was made to rule out various response tendencies in the 16 PF test, by balancing the number of "*Yes* or *a*" and "*No* or *c*" items scored for each scale (Cattell, 1956c). However, many of the items are not of the simple endorsement type, requiring instead a choice between alternative statements, e.g., "I would prefer the life of (*a*) an artist, (*b*) uncertain, (*c*) a secretary running a so-

TABLE 6

PERCENT "TRUE" AND "FALSE" ITEMS IN THE PRINCIPAL QUESTIONNAIRE SCALES LOADING E-I FACTORS

(Includes all items indicating agreement [True, Yes, Always, etc.] and disagreement [False, No, Never, etc.])

Scale	No. of items	% True	% False		No. of items	% True	% False
GZTS				16 PF			
G	30	63	37	(con'd.)			
R−	30	60	40	M−	13	23	31
A	30	53	47	Q1−	10	30	40
S	30	50	50	Q2−	10	30	30
T−	30	23	77	MMPI[c]			
STDCR				*L*−	15	100	0
S−	52	58	42	*D*−	60	67	33
R[a]	70	60	40	*Ma*	46	76	24
GAMIN				*Im*[d]	26	85	15
G	24	87.5	12.5	*Do*	28	25	75
A	38	61	39	*R*−	40	100	0
16 PF[b]				*Re*−	32	75	25
A	10	10	0	*Si*−	70	50	50
E	13	46	38.5	*Sp*[d]	27	56	44
F	13	46	15.5	*St*	34	44	56
H	13	38.5	38.5	*Sy*	32	69	31

Note.—To facilitate comparison, percentages for "introversion" scales have been reversed, as indicated by minus sign, e.g., R−. Thus, percentages indicate "True" and "False" items scored for *extraversion*.

[a] Items differentially weighted. Maximum score includes 53 points (62%) for *Yes*, 32 (38%) for *No* responses.

[b] Excluding "neutral" items (see text).

[c] Excluding scales with inconsistent factor loadings (e.g., *Hy*, *Pd*), and *Gm* and *M* (data not available).

[d] Unpublished keys; percentages based on data supplied by Harrison Gough.

cial club." Such items can scarcely reflect acquiescence, but, by the same token, they cannot be counted as part of the balanced distribution of items intended to rule it out. Thus, in determining the susceptibility of the 16 PF scales to acquiescence, these "neutral" items must be disregarded and consideration given only to the items which reflect agreement or disagreement. It can be seen in Table 6 that H and Q2 are equally weighted with "True" and "False" items; A, E, M, and Q1 differ by one item only. F, however, contains enough more "True" items so that scores on the F scale might be affected to some extent by acquiescence.

Turning to the MMPI scales, it can be seen that a "response set" interpretation of the factor obtained by Kassebaum, Couch, and Slater is supported not only by *R*, *L*, *Ma*, and *D*, but by the percentages of "True" and "False" items in the *Re*, *Sy*, *Sp*, and *Im* scales as well. Contrary to such an interpretation, however, are *Do* and *St*, which, though related to extraversion, contain more "False" items than "True" ones, and *Si*, whose items are evenly divided between the two categories.

It is quite possible that "psychological meaning" and acquiescence are confounded in a number of the questionnaire scales defining E-I factors, and until some means is found for distinguishing the two components, factor interpretations must take both aspects into account. At the same time, it is apparent that the E-I factors cannot be "explained" in terms of acquiescence alone. For the present, then, interpretations based on "psychological meaning" may be considered as having some validity.

Analyses of Objective Tests

Several objective test analyses related to E-I have appeared in recent years—some carried out by Cattell and his associates, others from Eysenck's laboratory. The latter studies rely heavily on tests of supposed or demonstrated relevance to particular dimensions of personality, whereas Cattell's analyses are based on tests intended to cover the entire "personality sphere." As might be expected, the test batteries used in the two sets of studies differ considerably, and the resulting E-I factors are not readily compared.

Analyses from Cattell's Laboratory

Cattell's objective test Factor UI 32, formerly Schizothyme Withdrawal, is now described as an extraversion factor (Cattell, 1957b); it has been renamed Exvia-Invia. One of the least confirmed objective test factors, UI 32 has appeared in only three analyses (CaA, CaB, Sc). As seen in Table 7, the loadings are generally small, and they vary somewhat from study to study. Nevertheless, there is some agreement as to the relative importance of fluency, ego strength, and inaccuracy—characteristics not infrequently associated with extraversion. A further link with E-I is provided by Cattell's CaB analysis, in which 16 PF Factors A, E, F, and H were found to correlate with UI 32. However, these findings are not supported by the more recent Scheier and Cattell analysis (Sc), in which only one of the questionnaire primaries—A—has a sizeable loading on UI 32. As Table 7 shows, F and H contribute little to the Scheier and Cattell E-I factor, and M, the highest-loading variable on Cattell's rating and questionnaire factors, has a zero loading; M appeared instead on a separate Autia factor (.40) and, negligibly, on UI 24, Anxiety vs. Dynamic Integration. M apparently failed to correlate with UI 32 in the CaB anal-

TABLE 7

OBJECTIVE TEST FACTORS: CATTELL'S LABORATORY

Variable	Factor Identification		
	CaA VII	CaB XV	Sc V-
Fluency on own characteristics	23	45	26
Fluency on self (vs. others) criticism			44
Fluency on dreams	13	21	
Rate of reading (delayed feedback)			−22A
Correctly articulated words (delayed feedback, reading)			−22
Correct word rate in reading (delayed feedback)			−36A
Immediate memory for words			−21
Myokinetic movement			−26
Objects perceived in unstructured drawings	09	49	—
Accuracy in gestalt completion	−14	−32	—
Ratio accuracy/accomplishment			−22A
Slanting line errors in CMS	−10	−31	29
Handwriting pressure			−31
High self-estimate of experience in various skills			20
Self-confidence in untried performance	13	30	10
Ego strength: Little shift to successfuls	20	03	30A
Authority suggestibility	−22	06	
Ratio acquaintances/friends			−22M
Preference for familiar (vs. strange) material			24
Preference for weak (vs. strong) smells			−24
Speed of regularly warned reaction time			−42
Pupil dilation at stress			−28
Increase in heart rate after startle			24
Systolic blood pressure			25
C: Free anxiety			28M
Q: 16 PF A, Cyclothymia		39[a]	53
E, Dominance		46[a]	
F, Surgency		46[a]	11
H, Parmia		43[a]	
M, Autia		—	01

Note.—Tests from The Objective-Analytic Personality Test Battery (Cattell, 1956a).
[a] Correlations between questionnaire scores and objective test factor.

ysis as well;[7] at least, no coefficient is given in a recent report of the study (Cattell, 1957b). In view of Cattell's insistence that "autia, M, belongs very definitely with the 'introversion' factors" (1957b, p. 317), his identification of UI 32 with E-I would seem to require clarification.

Analyses from Eysenck's Laboratory

Before turning to the objective test analyses carried out by Eysenck and his colleagues, it is necessary to say a few words about the underlying rationale. Eysenck's research over the years has culminated in a rather elaborate theory of extraversion-introversion (Eysenck, 1957)—essentially a rapprochment of the early views of Jung (1923) and McDougall (1926, 1929), Pavlov's concept of *inhibition*, and Hull's learning theory. Eysenck's theory has been criticized recently for its frequent failure to account for data it claims to explain (Storms & Sigal, 1958). It does, however, have much to recommend

[7] Although it had a substantial loading (.46) on UI 24, according to Cattell and Scheier (1958).

it, one of its chief assets being the ease with which it can be operationalized. Tests of the theory have been based for the most part on comparisons of two broadly defined groups of neurotics, believed to represent the extremes of the E-I continuum—*hysterics*, a group consisting of conversion hysterics and psychopaths, and *dysthymics*, a combination of anxiety neurotics, depressives, and obsessionals. The rationale for these groupings comes chiefly from Jung and McDougall, who regarded hysteria as the characteristic neurosis of extraverts, psychasthenia (anxiety, depression) as the typical introvert disorder. Eysenck added the remaining categories, and in an early factorial study (1944), obtained a "hysteria-dysthymia" factor which seemed to describe the two criterion groups.

It was noted above that analyses from Eysenck's laboratory have generally made use of tests selected for their relevance to particular personality dimensions. Since E-I has been a major area of interest for Eysenck and his co-workers, their analyses have generally included tests found—or hypothesized—to differentiate between hysterics and dysthymics. A number of such tests were included in Eysenck's first large-scale objective test study (1952), but while several factors emerged, none could be identified with E-I. Other analyses, however, have yielded factors which are at least suggestive of Eysenck's E-I dimension; these factors are shown in Table 8. The factors obtained by Heron (He) and Himmelweit, Desai, and Petrie (Him) have been discussed at length elsewhere (Eysenck, 1952, 1953) and require only brief mention. Appearing on these factors are a few tests found previously (Eysenck, 1947) to differentiate between hysterics and dysthymics—tests of persistence, and a couple of measures derived from level-of-aspiration experiments. Personal tempo loads one of the factors (Him), but Eysenck, in the publication just cited, has shown that his two criterion groups do not differ in this hypothetical E-I characteristic. Other tests supposedly related to E-I have negligible loadings on the factors; a few—fluency, quick approach to timed test (He), speed/accuracy ratio (Him)—have no loadings at all. In general, then, the relationship of the two factors to Eysenck's E-I dimension is not impressively demonstrated.

In a more recent analysis (Ey), Eysenck obtained an E-I factor defined by two sociometric measures of "sociability" and an index of performance speed—all theoretically related to extraversion, although the last one, at least, does not differentiate hysterics from dysthymics (Himmelweit, 1946). Apart from these measures, there is little to identify the factor with E-I. As Table 8 shows, the remaining E-I variables[8] have negligible loadings on the factor; others, hypothesized as measures of E-I, had essentially zero loadings: two tests of rigidity, a cognitive humor test, an affective discrepancy measure related to level-of-aspiration, and self-rated extraversion. The latter measure and teacher-rated extraversion, which has a loading of .18 on the factor, were based on rating scales adapted from Guilford's R scale—which, as noted previously, Eysenck regards as a good measure of his dimension! On the whole, then, the factor obtained in this analysis does not seem entirely consistent with E-I as Eysenck defines it.

The most impressive study coming

[8] Excluding projective test loadings, which appear in Table 9 and are discussed in conjunction with projective test analyses.

TABLE 8

OBJECTIVE TEST FACTORS: EYSENCK'S LABORATORY

Variable	Direction[a]	Factor Identification			
		Ey II[b,c]	He II	Hil III	Him II
Measures of E-I					
Porteus Mazes: Starting time	Quick			20	
Crossed lines	Many			43[d]	
Lifted pencils	Many	−11		01	
Wrong directions	Many	01		36[d]	
Track Tracer: Speed	High	38	−14		10A
Accuracy	Low	16			
Accuracy cost	Low		−32		
Personal tempo: Handwriting	Fast				34A
O'Connor Tweezer Test	Fast				26A
Rigidity: Alphabet test	Low	15			
Humor preference: Sex	High			22M	
Orectic humor	High	16			
Body build: Stature/transverse chest diameter	Short-round		−18	26	
Sociability I	High	63			
Sociability II	High	57			
I: Interests	Few		28M		
Q: STDCR S, Social Introversion	Low			47A[d]	
R, Rhathymia	High			56[d]	
R: Extraversion	High	18			
Measures of E-I and Neuroticism					
Persistence: Leg	Poor	−23	44	07	17M
Hand	Poor	−01	46		
Breath	Poor				22M
Level of aspiration: Mean goal discrepancy	Low positive				50
Absolute goal discrepancy	Low		−24		
Judgment discrepancy	High positive				17
Index of flexibility	High	−01			50
Measures of Neuroticism					
Crown Word Connection List: "Neurotic" score	High		33		
Track Tracer: Performance under stress	Poor				−14
O'Connor Tweezer Test: Evenness of improvement	Poor				52M
Body sway suggestibility: Total sway	High			05	13M
Reversals				26[d]	
Static ataxia: Total sway	High		−08M	18	11M
Reversals				07	
Dark vision	Poor				28
Systolic blood pressure[e]		−13			
Diastolic blood pressure[e]		−23			
Pulse rate after stress[e]		−12			
Sublingual temperature[e]		20			
Finger temperature[e]		30			
I: Annoyances	Many		−20M	−01M	
Q: Maudsley Medical Questionnaire (MMQ)	High			−26M	24
MMQ Lie Scale	High	−20[f]		−37A	
C: Mental health	Poor		30		
R: Neuroticism	High	17			
Unclassified Measures					
Perseveration: S-Z-SZ	Low		27		
237 and reversed	Low		28		
Strength of grip (hand dynamometer)	Strong		18		
I: Food aversions	Many		42		
Zygoticity[g]		22			

Note.—Variables with no loadings ≥.10 omitted from table; among them, several tests of E-I (see text) and intelligence.

[a] For *unclassified* measures, indicates scoring direction. For others, indicates predicted direction for *extraversion* (first two groups of variables) or *neuroticism* (third group). Variables reflected, when necessary, to agree with direction as listed here. Thus, positive loadings support prediction, negative ones do not.

[b] Projective test loadings for this factor appear in Table 9.

[c] No "adjustment" factor obtained in this analysis.

[d] Differentiates between 45 hysterics and 45 dysthymics at .05 or better (see text).

[e] Directional prediction not stated.

[f] Modified for children.

[g] Coded as follows: monozygotic, 1; dizygotic, 2.

from Eysenck's laboratory is the one reported recently by Hildebrand (Hil). Hildebrand tested 25 male normal subjects and a large group of male neurotics, including 45 hysterics (25 conversion hysterics, 20 psychopaths), 45 dysthymics (25 anxiety states, 10 depressives, 10 ob-

sessionals), and 55 cases with mixed symptomatology. In accord with Eysenck's theory, the conversion hysterics and anxiety states were reserved as E-I criterion groups; these two groups, together with the normal subjects, constituted criterion groups for neuroticism. A factor analysis was then carried out, using intercorrelations based on the remaining 95 subjects. Rotational criteria are not described in Hildebrand's article, but in a personal communication,[9] he indicates that Factor III, Extraversion-Introversion, was rotated to Guilford's R scale.

As Table 8 shows, all of the predicted E-I loadings on Hildebrand's factor are in the expected direction; some, however, are extremely small. Sizeable loadings on Guilford's R and S scales identify the factor with the previously discussed STDCR extraversion factors, but a question might still be raised about its relationship to Eysenck's dimension.

The question is answered by the second part of Hildebrand's analysis, in which E-I factor score comparisons were made for the various groups of subjects. Normal subjects were found to be the most extraverted, followed by hysterics, mixed neurotics, and dysthymics, in that order. Significant differences were obtained between conversion hysterics and anxiety states, and between the larger hysteric and dysthymic groups as well.

The results of these comparisons demonstrate convincingly the relationship of Hildebrand's E-I factor to Eysenck's dimension; they likewise seem to lend impressive support to Eysenck's theory. Not to be overlooked, however, are some important problems in the study itself and in the interpretation of the results. In the latter category, the greater extraversion of the normal group presents some difficulties, although, as Hildebrand suggests, it may simply reflect an unfortunate choice of control subjects. However, there is considerable evidence to indicate that conversion hysterics, at least, are no more extraverted than unselected normal subjects; they have consistently been found to score at or below the "normal" mean on E-I questionnaires (Eysenck, 1959; Sigal, Star, & Franks, 1958; Storms & Sigal, 1958). Calling attention to this finding, Eysenck (1959) notes that it "is not quite in line with expectation, but has been repeated on several samples and must be accepted" (p. 6).

Concerning the analysis itself, Storms and Sigal (1958) found from Hildebrand's original data (1953) that the groups pooled for the factor analysis differed significantly in variance on some of the tests. Discriminant functions computed for Hildebrand's data by Storms (1958) distinguished between conversion hysterics and anxiety states better than the factor scores, yet showed hysterics and psychopaths to be the most widely separated groups in terms of test performance. Hamilton (1957) has called attention to a similar lack of homogeneity within the dysthymic group, noting that on a dozen or so measures used in Hildebrand's study, either anxiety states or obsessionals performed more similarly to hysterics than to the other dysthymic subgroups. It is hard to tell whether these various inconsistencies can be attributed to the particular tests used, or whether they inhere in the criterion groups themselves. What evidence is available, however, seems to favor the latter explanation. In his early studies of hysterics and

[9] Hildebrand, H. P., Personal communication, March 4, 1959.

dysthymics, Eysenck (1947) repeatedly found larger standard deviations for the dysthymic group, leading him to suggest the possibility that "the dysthymic group contains several distinct subgroups" (p. 251). Moreover, data from a later study (Eysenck, 1952) showed no significant differences between hysterics, psychopaths, and anxiety states on tests of persistence, speed, accuracy, goal discrepancy, judgment discrepancy—all tests found previously to differentiate hysterics from dysthymics (Eysenck, 1947).

While these findings cast some doubt on the validity of the hysteric-dysthymic dichotomy, the fact remains that the two criterion groups *are* significantly differentiated by the major variables defining Hildebrand's factor (see Table 8) and by their E-I factor scores. Hildebrand's analysis thus establishes an important link between Eysenck's conception of E-I and the questionnaire factors defined by the Guilford scales.

Finally, brief mention should be made here of Becker's analysis (Be), discussed previously in connection with analyses of the Guilford and Cattell questionnaires. It will be recalled that Becker's E-I factor appeared to resemble the Lack of Self-Control factor obtained by Mann; its relationship to Eysenck's dimension is indicated by loadings of .81 and .82, respectively, on Guilford's R and Eysenck's E scale. The selection of objective tests for Becker's study was guided by Eysenck's recent theorizing about the relationship between E-I and "cortical inhibition" (1957). Among the tests were measures of reactive inhibition (reminiscence, response alternation), satiation (kinesthetic aftereffect, Archimedes spiral, Necker cube), and basal inhibition (GSR conditioning, aniseikonic lens tests, flicker fusion). Of the 32 variables derived from these tests, not one had a loading as great as .35 on the E-I factor. The only crucial variable to load over .30 was a kinesthetic aftereffect decrement score, which appeared to reflect little more than noncrucial differences in baseline, and which proved to have a retest reliability of zero. Thus, while the questionnaire loadings on Becker's factor readily identify it with Eysenck's concept of extraversion, it does not lend impressive support to more recent extensions of the concept.

Analyses of Projective Tests

The search for projective test counterparts of E-I has focused on the Rorschach test—the most widely studied projective instrument, and the only one linked by theory with the E-I dimension. It might be mentioned, however, that Sirota (1957) has identified an extraversion-like factor in another projective instrument—the psychoanalytically-oriented Blacky test (Blum, 1950). Sirota's factor, Impulse Expression vs. Impulse Control, may be related to the previously discussed "maladjusted extraversion" factors, but at present no empirical comparisons can be made.

The theoretical link between E-I and the Rorschach test is provided by Rorschach's concept of *experience balance*, expressed as the ratio of human movement (*M*) to color (*Sum C*) responses given to the Rorschach inkblots. *Extratensive* subjects, with a ratio favoring color, are said to be outwardly oriented, by virtue of their responsiveness to objective reality, i.e., color stimuli present in the blots. The perception of movement, on the other hand, has no corresponding external reality, and thus requires an intervening subjective process. Consequently, *introversive* subjects,

with a preponderance of movement responses, are described as having a more active "inner life" and less concern with external, objective reality.

While Rorschach (1951) denied any relationship between his experience balance concept and Jung's extraversion-introversion, the two viewpoints seem to have much in common. Rorschach's distinction between objective and subjective orientation is the crux of Jung's theory, and descriptions of the two Rorschach "experience types" are remarkably like Jung's characterizations of the extravert and introvert. Moreover, evidence from several studies indicates that some of the empirically found differences between extratensive and introversive subjects correspond to hypothesized or observed differences between extraverts and introverts (Bash, 1955; Bieri & Messerley, 1957; Mann, 1956; Palmer, 1957; Singer & Spohn, 1954).

Analyses of the Rorschach Test

Several Rorschach analyses have produced factors which appear to be related to experience balance, and which also have loadings on some non-Rorschach measures suggestive of E-I. The relevant factors are shown in Table 9, as are the projective test loadings for Eysenck's E-I factor (Ey), discussed above.

Eysenck included a number of Rorschach variables in his analysis, and obtained from a Rorschach "expert" opinions concerning their relevance to E-I. As Table 9 shows, Eysenck's extraversion has loadings on Rorschach *D*, *FM*:*M*, *F*%, and *P*; introversion is defined chiefly by *M*% and a composite pathology score. Expert opinion concurred with all but the *F*% loading, and Eysenck concludes that, on the whole, his results support the hypothesized relationship between E-I and Rorschach's extratension and introversion. It will be noted, however, that Eysenck's analysis included no color variables; his results thus say nothing about a relationship between extraversion and extratension. Nor can such a relationship be inferred from the fact that *M*% appears at the introvert pole, for Rorschach factors defined by *M* are not necessarily related to experience balance, as will be seen presently.

More pertinent to the experience balance question are two of the factors obtained in Singer, Wilensky, and McCraven's analysis (Si). Factor III, Emotional Surgency, combines positive loadings on the Rorschach color determinants with a small negative *M* loading; Factor IV, Introspectiveness, has substantial *positive* loadings on *M* and on a related measure, movement threshold (Barron, 1955). These factors, in turn, have loadings of .50 and −.32, respectively, on the first of two second-order factors reported by the authors, and thus seem to reflect a bipolar dimension of some sort. The movement-color contrast suggests that the dimension may be experience balance, and that Emotional Surgency and Introspectiveness correspond to Rorschach's extratension and introversion.

Concerning the relationship of the two factors to E-I, it might be noted that the ratings which appear on Factor III reflect a responsiveness to the environment (albeit a negative one!) which might suggest extraversion; likewise, acquiescence to authority and general disinterest in external events, associated with Factor IV, do not seem inconsistent with introversion. Moreover, the small level-of-aspiration loadings agree with Eysenck's (1947) findings for hysterics and dysthymics. Addi-

TABLE 9

PROJECTIVE TEST FACTORS

Variable	Factor Identification											
	Ey II[a]	FoA			FoB				Si			Wi III.[b]
		I	II	III	I	II-	IV	V	I	III	IV	
Barron Movement Threshold									56	01	62	
TAT Transcendence Index									−03	42	38	
Rorschach:												
W		63	−33	−07								
Wo					52	02	−47	−26				
Wv					00	43	−66	−02				
D	51											06, 48[c]
M		14	−45	75					81	−27	65	26, 36
FM					33	−56	−53	−02	−07	52	52	33, 39
Fv		50	22	−75								
C'									−19	57	−14	09, 26
FC									07	56	15	65, 48
CF									02	37	06	00, −12
C												20, −06
P	24											04, 35
R		71	28	34					07	53	44	27, 44
Rorschach ratios, %, etc.:												
F+%		14	−68	70					22	−08	14	
Special *F*+% (highly articulated responses)		18	−24	86								
Sum shading		60	21	05								
Sum C		69	50	00								
M : *Sum C*					36	−53	08	−09				
T: Wechsler-Bellevue												
Full Scale IQ									01	19	−12	
Verbal IQ		08	−81	06	−06	−51	00	20				00, −08
Vocabulary		44	−62	07	−11	−63	06	01				
Digit Span Scatter					−02	−47	−05	−20				
Wechsler Number-Square												
Initial performance									−18	18	−07	
Level of aspiration									23	−27	15	
Porteus Mazes									48	−07	−13	
Authority Reaction Test									04	11	57	
Motor inhibition									69	00	12	
Time estimation									17	28	02	
Digit Frustration									24	05	60	
C: Anxiety		65	24	−14								
R: Aggressiveness									−11	44	−06	
Cooperativeness									64	−53	−15	
Interest level									23	04	−59	
Diffuse energy level									−04	35	−21	
SR: Planfulness									02	07	−13	
Q: MMPI												
L, Lie Scale												−42, −20
F, Validity Scale												25, 12
K, Suppressor Scale												−39, −33
Hs, Hypochondriasis												−07, −25
D, Depression					72	−09	13	−43				−03, −40
Hy Hysteria												−48, −54
Pd, Psychopathic Deviate					52	25	13	44				−03, −09
Pa, Paranoia					22	−02	16	−07				07, 04
Pt, Psychasthenia					94	03	06	−03				14, 09
Sc, Schizophrenia					87	−01	16	31				22, 19
Ma, Hypomania					14	27	−18	72				59, 39
A, Anxiety												09, −17
R, Repression												−51, −55
Bernreuter												
F1-C, Confidence[d]					−80	24	−01	26				
F2-S, Sociability[d]					−24	66	15	00				
I: Allport-Vernon												
Theoretical					27	−64	16	22				
Economic					−37	10	−52	26				
Aesthetic					37	−27	42	−26				
Political					−10	67	−24	44				

Note.—Variables with no loadings ≥.10 omitted from table.

[a] Additional loadings: Rorschach *M*%, −63; *FM*: *M*, 50; composite pathological score, −40; *F*%, 29; *H*+*A*:*Hd*+*Ad*, −29; *Fm*+*m*:*Fc*+*c*+*C'*, −12; Rosenzweig Extrapunitiveness, 12.

[b] Additional loadings: Rorschach *m*, 47, 49; *d*, 28, 35; *F*, −20, 37; *Fc*, 17, 12; *K*, 16, −07; *c*, −01, 17.

[c] Orthogonal and oblique loadings, respectively.

[d] Bernreuter loadings reflected to agree with direction indicated by scale title.

tional evidence might be sought in Porteus Maze performance; the quantitative score used here is essentially an index of accuracy, and it ought to be closely related to the component variables—crossed lines,

wrong directions, etc.—previously associated with extraversion (Table 8). As Table 9 shows, however, the maze score contributes to neither factor above; it is found instead on Factor I, Motor Inhibition. This factor, like Introspectiveness, is suggestive of introversion; yet, despite some important similarities, the two factors are negatively correlated (−.15). Moreover, Motor Inhibition appears on the other second-order factor, which has zero loadings for Emotional Surgency and Introspectiveness. It would seem, then, that in addition to a pair of factors corresponding to experience balance, Singer et al. have uncovered a second, independent "introversion" factor in the Rorschach test. The latter factor, though unrelated to experience balance, seems as reasonable a match for E-I.

Singer et al. consider the three factors just discussed to be similar to Thurstone's Reflectiveness and Impulsivity factors (Table 3), and to the Emotional Drive and Inhibition factors obtained in two analyses by Foster (Table 9). Thurstone's factors can be related to Singer's only by inference, but in the case of Foster's factors, some direct comparisons can be made.

Factor I in Foster's first analysis (FoA) is called Emotional Drive; it shares with Singer's Surgency high loadings on Rorschach *R* and the color determinants (*Sum C* here). Factor III, Delay and Inhibition, appears most similar to Singer's Introspectiveness, although an important discrepancy is seen in the *F+* loadings.

In his second study, Foster used a modified Rorschach procedure to control for differential responsivity. Subjects were instructed to give at least three responses to each of the first nine cards, and at least six to Card X; the analysis was based on the required minimum (33 responses) for each subject. As can be seen in Table 9, the factors obtained in this analysis (FoB) are quite unlike those in the first study. Factor I, Delay and Inhibition, has a small *M*:*Sum C* loading, but the unusually high loadings on MMPI *Pt* and *Sc*—and on the Bernreuter Fl-C scale—mark it as a probable maladjustment factor. Factor IV, Emotional Drive, resembles its FoA counterpart in *W*, although the absence of the color component here argues against the identity of the two factors.

It is curious that in the search for factors comparable to their own, Singer and his associates overlooked the second factor in Foster's two analyses. These two factors have important loadings on Vocabulary and Verbal IQ, and Foster describes them as Verbal Adjustment factors. However, the prominent loadings on Rorschach *M* and *Sum C* (or *M*:*Sum C*) indicate that the two factors are closely allied with the experience balance concept; they seem to parallel the second-order experience balance factor obtained by Singer, Wilensky, and McCraven (1956). Some other variables appearing on the two factors (particularly the FoB factor) suggest a relationship to E-I; Bernreuter F2-S, MMPI *Ma*, perhaps the Allport-Vernon Political scale, which seems to be linked with extraversion (Eysenck, 1954). Even the strong verbal component might be looked upon as favorable evidence; Himmelweit (1945) has shown that dysthymics do better on vocabulary tests than on nonverbal measures of intelligence, whereas the reverse is true for hysterics.

The foregoing studies seem to support the validity of the experience balance concept, and they at least hint at a relationship between this

concept and E-I. On both points, however, there is equally impressive evidence to the contrary. Several extensive Rorschach analyses have failed to produce anything resembling an experience balance factor (Borgatta & Eschenbach, 1955; Karson & Pool, 1957a; Wittenborn, 1950a, 1950b), and inasmuch as these analyses do not seem to differ from the preceding ones in any consistent way, the discrepant results leave some doubt about the dimensionality of experience balance. In a similar vein, Williams and Lawrence (1953) obtained *two* factors with small contrasting loadings on *M* and *C;* in both cases, *FC* appeared at the "introvert" (i.e., *M*) extreme, and *CF* had no loading at all. Even the appearance in Singer, Wilensky, and McCraven's analysis of an "introversion" factor unrelated to experience balance muddies the waters considerably, although it does not, of course, rule out a relationship between E-I and experience balance. Finally, there is the evidence from comparisons of Rorschach measures with the various multidimensional questionnaires, below.

Joint Analyses: Rorschach and Questionnaire Measures

In an early attempt to demonstrate a relationship between E-I and experience balance, Thornton and Guilford (Tho) correlated various measures from the Rorschach test with Guilford's factors S, E, M, R, and T. They obtained no significant correlations between these factors and Rorschach *M*, *M*%, *C*, *C*%, or log *M*/*C*. Later, Royal (Ro) undertook a similar task, using the S, T, and R scales from the STDCR inventory. He was unable to find a single significant correlation between the three scales and a dozen potential Rorschach indices of E-I, including *M*, *Sum C*, and *M*: *Sum C*.

Similar results have been obtained with the MMPI. Palmer (1956) reports that chi square comparisons of the MMPI scores for 30 extratensive and 30 introversive subjects indicated no relationship between experience type and *Si* scores; other differences were "so few as to be of doubtful significance" (p. 208). Williams and Lawrence in a joint factor analysis (Wi) of the Rorschach and MMPI, obtained an "expressive-repressive" factor, shown in Table 9. The MMPI loadings on this factor are reminiscent of the E-I factor obtained by Kassebaum, Couch, and Slater. The Rorschach loadings, however, certainly do not correspond to experience balance. Foster's second analysis (FoB) is also relevant here. Of the three factors discussed above, only Factor I, Delay and Inhibition, has any sizeable MMPI loadings. The appearance of MMPI *D* on the factor is consistent with introversion, but equally so with maladjustment—a more reasonable interpretation in terms of the other loadings. Factor V in Foster's analysis, (Hypo)Manic-Depression, has not been mentioned previously. This factor, shown in Table 9, resembles the MMPI E-I factors considered earlier, but it has no important Rorschach loadings.

The projective test studies reviewed here attest to the unreliability of "apparent similarity" as a basis for matching factors. Some of the Rorschach factors appear to reflect certain characteristics associated with E-I; the identification is strengthened by occasional small loadings on E-I variables from other media. None of the evidence is very impressive, however, and the results of the joint analyses just discussed indicate that E-I questionnaire factors, at least, have little in common with the extraversion-like factors obtained from the Rorschach test.

Evaluation

To what extent has the nature of extraversion-introversion been clarified by recent multivariate research? What more—if anything—can be said about the unidimensionality of the construct, or its relationship to adjustment, on the basis of this research? These questions can perhaps be answered best by summing up the evidence in terms of the criteria set forth at the outset.

Extraversion-Introversion and Unidimensionality

The foregoing analyses indicate that it is possible to identify in all extensively studied measures and media at least one factor which bears some resemblance to traditional conceptions of E-I. The favorable results of early rating studies find confirmation in Cattell's discovery of an E-I factor in data from behavioral observation. Clear-cut factors have likewise emerged from analyses of various multidimensional questionnaires. Objective test batteries have in most cases yielded factors suggestive of E-I; in general, however, the factor loadings have been small, and interpretations somewhat uncertain. In the realm of projective tests, an extraversion-like factor has been found in the Blacky test, and factors identifiable with Rorschach's experience balance have appeared sporadically in analyses of the Rorschach test; the latter factors are linked by theory, at least, with E-I. In the various media, then, the situation remains essentially as Eysenck found it in 1953, with well-defined E-I factors appearing in questionnaire and rating studies, suggestive ones in analyses of objective and projective tests. True, a great deal more evidence has accumulated, particularly in the questionnaire medium, and much of it is favorable. Nevertheless, in terms of the first criterion—*the consistent appearance of E-I factors in all media of observation*—the unidimensionality of extraversion-introversion has not been conclusively demonstrated.

In terms of the second criterion—*the interrelatedness of the obtained factors*—the evidence is meager. No empirical comparisons have been reported for the objective or projective test factors obtained by different investigators; similarities have been noted in some cases, but the diversity of the variables, procedures, and populations represented in these studies makes speculation hazardous. Evidence from questionnaire studies shows that, in general, repeated analyses of the same instrument yield similar-appearing factors which, on the basis of "psychological meaning," can be identified with E-I. Such factors have been found in the questionnaires of Guilford and Cattell, and in the MMPI. Factor loadings vary from study to study, and variables are sometimes added or dropped, but there remains in each of the questionnaires a "core" of variables which appear consistently on E-I factors, regardless of the population studied, or the factorial procedure employed. Moreover, evidence from several studies shows that the core variables from the various questionnaires are at least moderately interrelated. Weighing against these very favorable findings, however, are the results of several joint analyses of the Guilford and Cattell questionnaires, showing that at least *two* independent factors are required to account for the intercorrelations between the E-I variables.

Little information is available concerning the relationships between E-I factors from different media. Cattell's rating and questionnaire factors appear similar, and a few of the questionnaire variables are re-

lated to his objective test E-I factor, although inconsistently. Objective test factors from Eysenck's laboratory are linked by one study with the Guilford questionnaire factors, by another—though less certainly—with some of the Rorschach variables. On the other hand, joint analyses of the Rorschach test and various questionnaires suggest that the extraversion-like factors from these instruments are probably unrelated.

It appears, then, that despite an impressive accumulation of relevant multivariate research, the unidimensionality of extraversion-introversion has not been unequivocally demonstrated.

Extraversion-Introversion and Adjustment

Except for the projective test analyses, where what constitutes an "adjustment" factor is not readily ascertained, virtually every analysis which has produced an extraversion-like factor has also yielded a factor identifiable with some aspect of adjustment. The latter factors, known variously as ego strength, general adjustment, neuroticism, anxiety, etc., appear to be essentially independent of E-I. The independence resulting from orthogonal rotation, while itself not impressive, tends to be supported by the few analyses employing oblique rotation. Cattell's second-order questionnaire factors of E-I and anxiety, for example, correlate −.02. Thus, according to the criterion of *uncorrelated factors*, extraversion-introversion and adjustment appear to be independent.

In many cases, however, it has been noted that the E-I factors seem to incorporate elements of adjustment. A glance at the factor patterns (Tables 2, 4, 7, 8) shows that most E-I factors share at least a few variables with adjustment factors from the same analyses. Looking at the questionnaire factors, it can be seen further that in analyses which have yielded a *single* E-I factor, the shared variables tend to align with that factor in such a way that "good" adjustment is associated with extraversion, "poor" adjustment with introversion. The tendency is most apparent in analyses of the Guilford and Cattell questionnaires (Table 2); in only one instance (Factor C in Analysis De) is an important *extra*version variable linked with *mal*adjustment. As might be expected, the tendency is less pronounced in the case of the MMPI factors (Table 4), where many of the variables are intrinsically related to maladjustment. Nevertheless, only two conspicuous exceptions are found —the *Ma* and *Pd* scales, which tend to be related to both extraversion and maladjustment. It is doubtful whether *Pd* should be counted; the *Pd* loadings on E-I factors are somewhat inconsistent. And while the *Ma* scale does appear consistently at the "extravert" extreme, there is some evidence that it may be related only to *mal*adjusted extraversion. In the case of questionnaire analyses yielding more than one extraversion-like factor, there are some indications that adjustment may be involved in the split. It has been noted in connection with these analyses that one of the factors generally bears some resemblance to "well-adjusted" extraversion, while another appears to reflect maladjusted extraversion. It has also been noted that such pairs of factors share few E-I variables, and thus seem to represent qualitatively different dimensions.

Turning to the factors from other media, it should be mentioned that none of the variables defining Cattell's E-I rating factor have loadings

as great as .30 on his second-order anxiety factor. However, the extraversion factor has a substantial negative loading on M, Autia, and if this primary rating factor is a true counterpart of questionnaire factor M, there is reason to suspect that introversion and maladjustment may be confounded in the E-I rating factor. By the same token, the apparent absence of M from Cattell's objective test factor UI 32 favors the independence of the latter factor from adjustment. Indeed, it can be seen in Table 7 that the few "adjustment" variables which appear on UI 32 are about evenly divided between the two poles of the factor. Unfortunately, the absence of M also raises a question about the relationship of UI 32 to Cattell's E-I factors in other media. Among the objective test analyses represented in Table 8, Eysenck's study yielded no adjustment factor, but his E-I factor links introversion with "pathology" as reflected in the Rorschach test. The remaining factors shown in Table 8 are similar to Cattell's UI 32 in the division of "adjustment" variables. As was the case with UI 32, however, the identification of some of these factors with E-I might be questioned.

If it is asked, then, whether extraversion-introversion and adjustment are independent, in the sense that *variables reflecting "good" and "poor" adjustment are as frequently associated with extraversion as with introversion*, a clear-cut answer cannot be given. It is evident that many of the questionnaire factors do not meet this second criterion, and for most of the factors which *are* independent in this sense, there is some doubt about their relationship to E-I.

Concluding Remarks

The present review was prompted by the recent burgeoning of interest in extraversion-introversion, and by the fact that current assumptions about the unidimensionality of the construct, and its independence from adjustment, cannot be justified in terms of the research covered by the last comprehensive review (Eysenck, 1953). An examination of more recent research has shown the evidence on both issues to be equivocal, and the status of extraversion-introversion as a dimension of personality thus remains somewhat tenuous.

In concluding, it is well to point out what appear to be the major implications of the research reviewed here. First, the "nomological network" developing from Eysenck's earlier review has begun to be tied down to observable data—a replicated factor here, a series of intercorrelations there—and, while a great many gaps remain, there is reason to believe that further research along these lines will not be wasted. Second, the most profitable directions for such research seem to be clearly indicated. There are variables whose relationships to extraversion and adjustment need to be clarified. There are factors whose widely differing patterns across studies need to be accounted for. There are areas which have not been—or are just beginning to be—systematically explored. There are hints that extraversion-introversion may be differentially manifested in males and females, and in well-adjusted and maladjusted individuals; both possibilities need to be followed up. Finally, and perhaps most important, there is a need for broadly conceived analyses oriented toward extraversion-introversion and its relationship to adjustment. Such analyses would necessarily include a wide array of variables from all media—variables

selected for their relevance to the two dimensions, and, when possible, variables of known factorial composition, so that the resulting factors could be compared empirically with previously discovered ones. Until such further steps are taken, the issues raised here are not likely to be resolved.

In the meantime, a word of caution seems in order. If the term *extraversion-introversion* is to continue in psychological usage—and, judging from past history, there is little likelihood that it will not—care must be taken to specify its conceptual and operational referent. What appear to be minor distinctions between the various conceptions may in fact be crucial ones; to discard them too hastily is likely only to propagate the illusion of a unity not yet established.

REFERENCES

ABRAMS, E. N. A comparative factor analytic study of normal and neurotic veterans. Unpublished doctoral dissertation, Univer. of Michigan, 1949.

BAEHR, MELANY E. A factorial study of temperament. *Psychometrika*, 1952, **17**, 107–126.

BARRON, F. Threshold for the perception of human movement in ink-blots. *J. consult. Psychol.*, 1955, **19**, 33–38.

BASH, K. W. Einstellungstypus and Erlebnistypus: C. G. Jung and Hermann Rorschach. *J. proj. Tech.*, 1955, **19**, 236–242.

BECKER, W. C. Cortical inhibition and extraversion-introversion. USPHS Proj. No. M1713, Univer. Illinois, 1959.

BERNREUTER, R. G. The imbrication of tests of introversion-extroversion and neurotic tendency. *J. soc. Psychol.*, 1934, **5**, 184–201.

BIERI, J., & MESSERLEY, SUSAN. Differences in perceptual and cognitive behavior as a function of experience type. *J. consult. Psychol.*, 1957, **21**, 217–221.

BLUM, G. S. *The Blacky pictures: A technique for the exploration of personality dynamics.* New York: Psychol. Corp., 1950.

BORGATTA, E. F., & ESCHENBACH, A. E. Factor analysis of Rorschach variables and behavioral observation. *Psychol. Rep.*, 1955, **1**, 129–136.

CATTELL, R. B. The diagnosis and classification of neurotic states: A reinterpretation of Eysenck's factors. *J. nerv. ment. Dis.*, 1945, **102**, 567–589.

CATTELL, R. B. *The description and measurement of personality.* Yonkers, N. Y.: World Book, 1946.

CATTELL, R. B. *Personality.* New York: McGraw-Hill, 1950.

CATTELL, R. B. Psychiatric screening of flying personnel: I. Personality structure in objective tests. *USAF Sch. Aviat. Med. Rep.*, 1955, No. 9.

CATTELL, R. B. *The Objective-Analytic Personality Test Battery.* Champaign, Ill.: Inst. Pers. Abil. Testing, 1956. (a)

CATTELL, R. B. Second-order factors in the questionnaire realm. *J. consult. Psychol.*, 1956, **20**, 411–418. (b)

CATTELL, R. B. Validation and intensification of the Sixteen Personality Factor Questionnaire. *J. clin. Psychol.*, 1956, **12**, 205–214. (c)

CATTELL, R. B. *Handbook for the Sixteen Personality Factor Questionnaire.* Champaign, Ill.: Inst. Pers. Abil. Testing, 1957. (a)

CATTELL, R. B. *Personality and motivation structure and measurement.* Yonkers, N. Y.: World Book, 1957. (b)

CATTELL, R. B., & SCHEIER, I. H. The nature of anxiety: A review of thirteen multivariate studies comprising 814 variables. *Psychol. Rep.*, 1958, **4**, 351–388.

COOK, E. B., & WHERRY, R. J. A factor analysis of MMPI and attitude test data. *J. appl. Psychol.*, 1950, **34**, 260–266.

COTTLE, W. C. A factorial study of the Multiphasic, Strong, Kuder, and Bell inventories using a population of adult males. *Psychometrika*, 1950, **15**, 25–47.

DENTON, J. C., & TAYLOR, C. W. A factor analysis of mental abilities and personality traits. *Psychometrika*, 1955, **20**, 75–81.

ERIKSEN, C. W. Psychological defenses and "ego strength" in the recall of completed and incompleted tasks. *J. abnorm. soc. Psychol.*, 1954, **49**, 45–50. (a)

ERIKSEN, C. W. Some personality correlates of stimulus generalization under stress. *J. abnorm. soc. Psychol.*, 1954, **49**, 561–565. (b)

ERIKSEN, C. W., & DAVIDS, A. The meaning and clinical validity of the Taylor anxiety scale and the hysteria-psychasthenia scales from the MMPI. *J. abnorm. soc. Psychol.*, 1955, **50**, 135–137.

EYSENCK, H. J. Types of personality: A factorial study of seven hundred neurotics. *J. ment. Sci.*, 1944, **90**, 851–861.

EYSENCK, H. J. *Dimensions of personality.* London: Routledge & Kegan Paul, 1947.

EYSENCK, H. J. *The scientific study of personality.* London: Routledge & Kegan Paul, 1952.

EYSENCK, H. J. *The structure of human personality.* London: Methuen, 1953.

EYSENCK, H. J. *The psychology of politics.* London: Routledge & Kegan Paul, 1954.

EYSENCK, H. J. The inheritance of extraversion-introversion. *Acta psychol. Amsterdam,* 1956, **12,** 95–100 (a)

EYSENCK, H. J. The questionnaire measurement of neuroticism and extraversion. *Riv. psicol.,* 1956, **50,** 113–140. (b)

EYSENCK, H. J. *The dynamics of anxiety and hysteria.* New York: Praeger, 1957.

EYSENCK, H. J. *Manual of the Maudsley Personality Inventory.* London: Univer. London Press, 1959.

FOSTER, A. The factorial structure of the Rorschach test. *Texas Rep. Biol. Med.,* 1955, **13,** 34–61.

FRANKS, C. M. Conditioning and personality: A study of normal and neurotic subjects. *J. abnorm. soc. Psychol.,* 1956, **52,** 143–150.

FRANKS, C. M., & LAVERTY, S. G. Sodium amytal and eyelid conditioning. *J. ment. Sci.,* 1955, **101,** 654–663.

FREUD, S. *A general introduction to psychoanalysis.* New York: Liveright, 1920.

GUILFORD, J. P. Introversion-extroversion. *Psychol. Bull.,* 1934, **31,** 331–354.

GUILFORD, J. P. *An Inventory of Factors STDCR.* Beverly Hills, Calif.: Sheridan Supply Co., 1940.

GUILFORD, J. P., & GUILFORD, RUTH B. An analysis of the factors in a typical test of introversion-extroversion. *J. abnorm. soc. Psychol.,* 1934, **28,** 377–399.

GUILFORD, J. P., & GUILFORD, RUTH B. Personality factors S, E., and M, and their measurement. *J. Psychol.,* 1936, **2,** 109–127.

GUILFORD, J. P., & GUILFORD, RUTH B. Personality factors D, R, T, and A. *J. abnorm. soc. Psychol.,* 1939, **34,** 21–36. (a)

GUILFORD, J. P., & GUILFORD, RUTH B. Personality factors N and GD. *J. abnorm. soc. Psychol.,* 1939, **34,** 239–248. (b)

GUILFORD, J. P., & HUNT, J. M. Some further experimental tests of McDougall's theory of introversion-extroversion. *J. abnorm. soc. Psychol.,* 1932, **26,** 324–332.

GUILFORD, J. P., & MARTIN, H. *The Guilford-Martin Inventory of Factors GAMIN.* Beverly Hills, Calif.: Sheridan Supply Co., 1943. (a)

GUILFORD, J. P., & MARTIN, H. *The Guilford-Martin Personnel Inventory I.* Beverly Hills, Calif.: Sheridan Supply Co., 1943. (b)

GUILFORD, J. P., & ZIMMERMAN, W. S. *The Guilford-Zimmerman Temperament Survey.* Beverly Hills, Calif.: Sheridan Supply Co., 1949.

GUILFORD, J. P., & ZIMMERMAN, W. S. Fourteen dimensions of temperament. *Psychol. Monogr.,* 1956, **70**(10, Whole No. 417).

HAMILTON, V. Conflict avoidance in obsessionals and hysterics, and the validity of the concept of dysthymia. *J. ment. Sci.,* 1957, **103,** 666–676.

HATHAWAY, S. R., & MCKINLEY, J. C. *Manual for the Minnesota Multiphasic Personality Inventory.* New York: Psychol. Corp., 1951.

HERON, A. The objective assessment of personality among factory workers. *J. soc. Psychol.,* 1954, **39,** 161–185.

HILDEBRAND, H. P. A factorial study of introversion-extraversion by means of objective tests. Unpublished doctoral dissertation, Univer. of London, 1953.

HILDEBRAND, H. P. A factorial study of introversion-extraversion. *Brit. J. Psychol.,* 1958, **49,** 1–11.

HIMMELWEIT, H. T. The intelligence-vocabulary ratio as a measure of temperament. *J. Pers.,* 1945, **14,** 93–105.

HIMMELWEIT, H. T. Speed and accuracy of work as related to temperament. *Brit. J. Psychol.,* 1946, **36,** 132–144.

HIMMELWEIT, H. T., DESAI, M., & PETRIE, A. An experimental investigation of neuroticism. *J. Pers.,* 1946, **15,** 173–196.

HOVEY, H. B. Measures of extraversion-introversion tendencies and their relation to performance under distraction. *J. genet. Psychol.,* 1929, **36,** 316–329.

JUNG, C. G. *Psychological types.* (Trans. by H. G. Baynes) New York: Harcourt, Brace, 1923.

KARSON, S. Second-order personality factors and the MMPI. *J. clin. Psychol.,* 1958, **14,** 313–315.

KARSON, S., & POOL, K. B. The abstract thinking abilities of mental patients. *J. clin. Psychol.,* 1957, **13,** 126–132. (a)

KARSON, S., & POOL, K. B. The construct validity of the Sixteen Personality Factors Test. *J. clin. Psychol.,* 1957, **13,** 245–252. (b)

KARSON, S., & POOL, K. B. Second-order factors in personality measurement. *J. consult. Psychol.,* 1958, **22,** 299–303.

KASSEBAUM, G., COUCH, A. S., & SLATER, P. E. Factorial dimensions of the MMPI. *J. consult. Psychol.,* 1959, **23,** 226–236.

LAVERTY, S. G., & FRANKS, C. M. Sodium amytal and behavior in neurotic subjects. *J. Neurol. Neurosurg. Psychiat.,* 1956, **19,** 137–143.

LOVELL, CONSTANCE. A study of the factor structure of thirteen personality variables. *Educ. psychol. Measmt.,* 1945, **5,** 335–350.

McDOUGALL, W. *Outline of abnormal psychology.* New York: Scribner's, 1926.

McDOUGALL, W. The chemical theory of temperament applied to introversion and extroversion. *J. abnorm. soc. Psychol.*, 1929, **24**, 293–309.

MANN, L. The relation of Rorschach indices of extratension and introversion to a measure of responsiveness to the immediate environment. *J. consult. Psychol.*, 1956, **20**, 114–118.

MANN, R. D. The relationship between personality characteristics and individual performance in small groups. Unpublished doctoral dissertation, Univer. of Michigan, 1958.

MOORE, H., & STEELE, ISABEL. Personality tests. *J. abnorm. soc. Psychol.*, 1934, **29**, 45–52.

NELSON, M. O., & SHEA, SALLY. MMPI correlates of the inventory of factors STDCR. *Psychol. Rep.*, 1956, **2**, 433–436.

NORTH, R. D. An analysis of the personality dimensions of introversion-extroversion. *J. Pers.*, 1949, **17**, 352–367.

PALMER, J. O. Attitudinal correlates of Rorschach's experience balance. *J. proj. Tech.*, 1956, **20**, 207–211.

PALMER, J. O. Some relationships between Rorschach's experience balance and Rosenzweig's frustration-aggression patterns. *J. proj. Tech.*, 1957, **21**, 137–141.

RORSCHACH, H. *Psychodiagnostics.* (5th ed., trans. by P. Lemkau & B. Kronenberg) Berne, Switzerland: Verlag Hans Huber, 1951.

ROYAL, R. E. An experimental investigation of the relationship between questionnaire and Rorschach measures of introversion. Unpublished doctoral dissertation, Univer. of Pittsburgh, 1950.

SCHEIER, I. H., & CATTELL, R. B. Confirmation of objective test factors and assessment of their relation to questionnaire factors: A factor analysis of 113 rating, questionnaire and objective test measurements of personality. *J. ment. Sci.*, 1958, **104**, 608–624.

SIGAL, J. J., STAR, K. H., & FRANKS, C. M. Hysterics and dysthymics as criterion groups in the study of introversion-extraversion. *J. abnorm. soc. Psychol.*, 1958, **57**, 143–148.

SINGER, J. L., & SPOHN, H. E. Some behavioral correlates of Rorschach's experience-type. *J. consult. Psychol.*, 1954, **18**, 1–9.

SINGER, J. L., WILENSKY, H., & McCRAVEN, VIVIAN G. Delaying capacity, fantasy, and planning ability: A factorial study of some basic ego functions. *J. consult. Psychol.*, 1956, **20**, 375–383.

SIROTA, L. M. A factor analysis of selected personality domains. Unpublished doctoral dissertation, Univer. of Michigan, 1957.

STAGNER, R. The intercorrelation of some standardized personality tests. *J. appl. Psychol.*, 1932, **16**, 453–464.

STORMS, L. H. Discrepancies between factor analysis and multivariate discrimination among groups as applied to personality theory. *J. ment. Sci.*, 1958, **104**, 713–721.

STORMS, L. H., & SIGAL, J. J. Eysenck's personality theory with special reference to "The dynamics of anxiety and hysteria." *Brit. J. med. Psychol.*, 1958, **31**, 228–246.

THORNTON, G. R., & GUILFORD, J. P. The reliability and meaning of Erlebnistypus scores in the Rorschach test. *J. abnorm. soc. Psychol.*, 1936, **31**, 324–330.

THURSTONE, L. L. The dimensions of temperament. *Psychometrika*, 1951, **16**, 11–20.

TYLER, F. T. A factorial analysis of fifteen MMPI scales. *J. consult. Psychol.*, 1951, **15**, 451–456.

VERNON, P. E. *The assessment of psychological qualities by verbal methods.* London: Her Majesty's Stationery Office, 1938.

WEITZENHOFFER, A. Hypnotic susceptibility as related to masculinity-feminity. Unpublished doctoral dissertation, Univer. of Michigan, 1956.

WELSH, G. S. Factor dimensions *A* and *R*. In G. S. Welsh & W. G. Dahlstrom (Ed.), *Basic readings on the MMPI in psychology and medicine.* Minneapolis: Univer. Minnesota Press, 1956. Pp. 264–281.

WHEELER, W. M., LITTLE, K. B., & LEHNER, G. F. J. The internal structure of the MMPI. *J. consult. Psychol.*, 1951, **15**, 134–141.

WILLIAMS, H. L., & LAWRENCE, J. F. Further investigation of Rorschach determinants subjected to factor analysis. *J. consult. Psychol.*, 1953, **17**, 261–264.

WILLIAMS, H. L., & LAWRENCE, J. F. Comparison of the Rorschach and MMPI by means of factor analysis. *J. consult. Psychol.*, 1954, **18**, 193–197.

WITTENBORN, J. R. A factor analysis of Rorschach scoring categories. *J. consult. Psychol.*, 1950, **14**, 261–267. (a)

WITTENBORN, J. R. Level of mental health as a factor in the implications of Rorschach scores. *J. consult. Psychol.*, 1950, **14**, 469–472. (b)

WOOD, R. L. Prediction and analysis of attrition in classes of a university reading service. Unpublished doctoral dissertation, Univer. of Michigan, 1957.

(Received July 20, 1959)

Psychological Bulletin
1960, Vol. 57, No. 5, 361–402

COMPARATIVE PSYCHOLOGICAL STUDIES OF NEGROES AND WHITES IN THE UNITED STATES[1]

RALPH MASON DREGER
Jacksonville University
AND KENT S. MILLER
Leon County Health Department, Tallahassee, Florida

This report is a review of published psychological studies which involve a comparison of Negroes and whites in the United States during the period in the main from 1943–1958. The period chosen covers work reported following Klineberg's (1944) review. Where called for, for completeness, we have selected studies from previous years. Most of the literature to which reference is made reports direct comparisons between Negroes and white groups. However, in some instances direct comparisons were not made, but inferences could be drawn —e.g., where a test standardized on a white population has been administered to a Negro group. For the purpose of inclusion in this review, research has been considered relevant wherever the individual authors have stated that Negro-white comparisons were made or wherever the stated population in studies utilizing white-standardized tests is Negro. Thus the populations compared have consisted of varying degrees of "racial pureness." We have reached into studies of physical development and sociological research where we have deemed it important for psychological completeness.

Two earlier reviews of racial psychology (primarily concerned with mental differences) have appeared in this Journal (Garth, 1925; Woodworth, 1916). For comprehensive discussions of the Negro and his problems the reader is referred to Myrdal (1944), Klineberg (1944) and Canady (1946).

In addition to these references, racial differences are discussed in many general works, some prior to our period (Anastasi & Foley, 1949; Bendix & Lipset, 1953; Benedict & Weltfish, 1943; Boyd, 1950; Dunn & Dobzhansky, 1946; Frazier, 1957; Garth, 1931; Ginzberg, 1956; Kardiner & Ovesey, 1951; Knox, 1945, 1949, 1952; Lindzey, 1954; Montagu, 1952; Sarason & Gladwin, 1958; Tyler, 1956). We have endeavored in the following review to confine ourselves to the experimental literature; only in the sections on values and attitudes and on emotional disturbances have we departed to some extent from the rule. Over 200 additional references were reviewed, but not included among our references, either because they have been adequately covered in other reviews, or they lacked significance or relevancy. Less than a dozen possibly significant theses were abstracted in various places, but were not studied by the writers.

Physical and Motor Development

Physical Status of Infants and Children

Gestation, birth, and early physical development may have decisive influence upon later psychological functioning but are of importance in themselves. Even later infancy is measured psychologically in terms of fine and gross muscle coordina-

[1] Prepared with the assistance of a grant from the Human Genetics Fund.

tions. Accordingly, we summarize briefly studies of physical and motor development of children.

Brown, Lyon, and Anderson (1945) report a significantly higher rate of prematurity among Negro infants than among whites if criterion of birth weight is used as a standard. However, if allowance is made for the generally known fact that Negro infants weigh less at birth (for whatever reasons), the difference in prematurity rates disappears.

White infants are on the average longer than Negro infants if the results of a 1943 survey (Meredith, 1943) of North American research still hold. The possibility is that some change occurs as socioeconomic differences are reduced. At least, such an inference could be made from studies conducted in which socioeconomic levels have been held constant (Rhoads, Rapoport, Kennedy, & Stokes, 1941; Scott, Cardozo, Smith, & DeLilly, 1950). Rhoads et al. (1941) followed white and Negro children from four months to four years in the outpatient department of a children's hospital. From the same lower socioeconomic groups, Negro infants actually tended to be taller from about nine months of age. Body weight and head circumference, eruption of teeth and motor development of both groups were about the same. Negro children tended to have longer legs, shorter trunks, smaller body width at chest and pelvis, lower hemoglobin count, and better teeth. No differences in height and weight were observed by Scott et al. when measurements of 654 Negro infants in the first year of life from the lower middle class were compared with those of white infants from comparable socioeconomic levels. Michelson (1943) concluded that the weight patterns are very similar when individuals are placed on the same dietary regime.

Psychomotor Development

Tests of infant intelligence have been suspect (possibly unjustifiably) for many years as predictors of subsequent intellectual performance. Nevertheless, as measures of comparative psychomotor developmental level, they perform a useful task.

The most widely known study is that of McGraw (1931) which was reported prior to the period covered by this review. In a comparison of Negro and white infants in a Southern community on the Buhler Baby Tests, McGraw discovered that white babies exceeded Negro babies in performance. The former were also superior in height and weight.

Pasamanick (1946), contending that McGraw was measuring differences in opportunities for adequate care and feeding rather than native abilities, endeavored to overcome objections to the previous work by utilizing New Haven infants somewhat comparable in weight and height factors. Comparison on an infant development scale suggested that differences are not between groups so much as within racial groups. Negro babies proved to be somewhat accelerated relatively in gross motor behavior.

Inadequacies of sampling, as well as the more general difficulty of estimating skin color subjectively, tend to invalidate Pasamanick's conclusions. His infants may very well have been comparable for other reasons than the quality of diet during and after pregnancy to which Pasamanick attributes his results.

In a follow-up study utilizing the same Negro infants Pasamanick and Knoblock (1955) determined that depressing influences on the Develop-

mental Quotient of exogenous factors were not important up to at least two and one half years of age. Whatever methodological difficulties are suggested by the original study, the fact that the group of Negro infants maintained their DQs for so long a period of time indicates that they were maintaining their relative developmental standing, including intellectual functioning.

What appears to be better-controlled research by Gilliland (1951) employed the Northwestern Infant Intelligence Scale rather than the Gesell scales (understandably, since Gilliland developed the former). In three separate studies, at least one of which controlled for major variables, Negro infants in Chicago had IQs as high as or higher than white infants. The results may, the author suggests, be interpreted as manifestations of the greater maturity in motor behavior Negro babies are known to have, or as a resultant of crowded quarters in which Negro babies receive greater social stimulation. Inasmuch as the former explanation does not seem to be found when socioeconomic variables are controlled, the latter may be more acceptable.

The remainder of the comparisons of psychomotor and physical development is summarized in Table 1. Generally, the results indicate white and Negro children represent the same populations in respect to each of the variables measured.

In a study obliquely related to white-Negro comparisons, Codwell (1949) separated a Negro group into three groups varying in degrees of "Negroidness." Composite motor functioning did not change from one group to another, implying no differences between those more Negroid and those more white.

Overview of Physical and Motor Development

On some physical measurements of children, especially anatomical measurements of substructures, a racial difference appears to exist between whites and Negroes. After equating for socioeconomic variables, investigators on the whole find that differences in psychomotor functioning tend to disappear.

Psychophysical Functions

A few studies have endeavored to compare white and Negro subjects on variables which may roughly be

TABLE 1

Comparisons of Infants and Children in Psychomotor and Physical Development

Author(s)	Topic	Subjects	Results	Comments
Williams & Scott (1953)	Development (Gesell)	W low S-E[a] status, N high S-E status	W > N	Authors attribute to differential child-rearing, e.g., greater permissiveness in lower class
Scott, Ferguson, Jenkins, & Cutler (1955)	12 neuromuscular steps	Infants	N > W	N from low S-E class > N from higher S-E class up to 35th week, same following that age
Irwin (1949)	Speech	Infants: 1–10 days 1st 30 mos.	W ≅ N W ≅ N	Professionals' children have steeper parabolic curves of phonetic development than laborers' children
Moore (1942)	Eye-hand coordination	Preschool	W ≅ N	From 24–35 mos. N child superior to W
Espenschade (1946)	Coordination and agility	10th grade girls	W ≅ N	
Ramsey (1950)	Pubertal changes	Adolescent boys	W ≅ N	

[a] Socioeconomic.

classed as biological, but which have, we know, psychological components of varying degrees of importance.

Measurements of pain thresholds by Chapman (1944) suggest that Negroes perceive radiant heat at a lower threshold than do whites. In this experiment 18 Negroes and 18 Americans of southern European ancestry were employed; but even though for psychophysical experiments a small *N* would be satisfactory, the use of psychoneurotics as well as normal persons might bias the results for a normal population. Further, only raw data are given without appropriate statistical treatment.

Among selectees and inductees, 2,200 Negro and 21,000 whites, it was revealed that more colored selectees had normal vision in each eye than did whites. No decrease in the differences was noted with age. It was also determined that the poorer eye fared better among Negroes than the better one did among whites. Karpinos (1944) who reports the study gives no reason for the findings. In the Canal Zone, Covell (1950) discovered that Negroes develop presbyopia 5 to 10 years before whites do. Other than a higher incidence of lues and tuberculosis, the usually accepted causes of presbyopia do not seem to offer an explanation for the differences. Personal insecurity and social inequality are suggested by the author to be of fundamental importance in setting off organic changes within the lens. Peripheral absolute visual thresholds are lower under dark adaptation in Negroes and become progressively higher as the pigment of the fundus gets lighter or is reduced, according to a study by Helson and Guilford (1933). As the center of vision is approached, the difference grows less (as it does between white persons with different colored eyes). Woisika (1944) concludes that the Negro is superior to the white in dark adaptation. Correlation of age and dark adaptation (Negro mean age = 39, white = 46) may vitiate the conclusion.

Some racial differences appear in psychophysical functions. Much more research of the type Eysenck, Granger, and Brengelmann (1957) have done on perceptual processes and mental illness will have to be done in order to discover what part genetic factors play in these differences.

Intellectual Functions

Shuey (1958) has reviewed the literature comparing Negro and white intelligences at least as far back as 1913. We shall not endeavor to cover the same ground. In her text are found valuable tabular comparisons of Negroes and whites for various age groups, the armed forces, gifted and retarded, delinquents and criminals, and racial hybrids. (Actually a number of studies reported in the literature are of hybrids even though not recognized as such.) Shuey's bibliography and résumés are a must for serious students, for she has not only gathered together the better known studies, but has ferreted out otherwise obscure and inaccessible articles and theses.

The usefulness of Shuey's otherwise excellent work is limited by what appears to be a polemic attitude. Her book seems to be an attempt to prove a nonegalitarian hypothesis rather than being strictly a review of literature. In this case Shuey does the same rationalizing from an hereditarian standpoint that Klineberg (1944) did in his earlier "review" from an environmental standpoint.

North (1957) likewise surveys the literature and comes to an opposite conclusion from the one Shuey

reaches. Whereas the latter concluded that all the evidence points " . . . to the presence of some native differences between Negroes and whites as determined by intelligence tests," North maintains that there is no proof of biological inferiority or that the Negro's potentials for educational and cultural development are more limited than the white person's. North's coverage of the literature is considerably less extensive than Shuey's.

The following discussion on comparisons of intelligence between the two racial groups endeavors to supplement Shuey's work especially and correct it where it is in patent error.

Children and Adolescents

Young Children

Aside from the infant studies cited previously, studies of preschool children in which comparisons are made between whites and Negroes are sparse. Shuey (1958) cites only nine altogether and only five reported in 1944 or after. On the whole, young Negro children score lower than whites. But the differences are very much less than in older groups; and in all of the reports in which average IQs are given, Negroes average well within the normal IQ range for whites. Shuey offers several explanations of the discrepancy between the results for preschool and school children, including inadequate sampling of preschool, the relative invalidity and unreliability of tests at younger ages (although some of the lowest Binet standard errors are found in the late preschool years), more verbal and abstract tests in the school years; she also suggests that mental growth curves may not be the same for both races, and that IQs may be less affected by environment in the preschool years.

Special comment is called for with respect to two investigations. Brown (1944) compared Minneapolis kindergarten children on the Binet, Form L, discovering that the Negro mean of 100.8 was significantly lower than the white mean of 107.1. The white children in Occupational Classes VI and VII (Minnesota scale) averaged about the same as Negro children. Shuey takes exception to Brown's conclusion that at "nominally similar" socioeconomic status Negro children are not inferior to whites. Her observation that Brown has small *N*s in Levels VI and VII does not constitute an objection from a statistical standpoint. Nevertheless, Brown's conclusion would have been stronger if he had, first, differentiated his Negroes by class without assuming their occupational status and, second, employed an analysis of variance design.

The other study to which special attention must be paid is that of Anastasi and D'Angelo (Anastasi & D'Angelo, 1952; D'Angelo, 1950). Five-year-old children in mixed and unmixed neighborhoods in New York City were administered the Goodenough Draw-a-Man Test and studied for language development in spontaneous conversation recordings. IQs were 101.8 and 101.5 for Negro and white children, respectively. Language development appeared to be somewhat more advanced for white than for colored children. This important study is dismissed by Shuey as permitting no generalizations, because selection of subjects appeared to be biased. The crux of the matter is whether all qualified subjects were utilized or only certain selected ones. Anastasi[3] points out Shuey's error in misjudging the selection procedure, and provides a satisfactory

[3] Anastasi, Anne. Personal communication, July 19, 1958.

answer to Shuey's objections. With due recognition of the limitations of the Goodenough as a test of intelligence we may yet regard Anastasi and D'Angelo's results as a challenge to nativist theories of intellectual differences between the races.

Older Children and Adolescents

Little question need be raised at this time about the results of testing school children. Almost all evidence points to inferior performance of Negroes on tests of either the more traditional variety or those tending to be more "culture-free" or "culture-fair." What Shuey (1958) has done in this respect is not to present startlingly new conclusions, but to marshall data which have been more or less familiar to scholars for many years. Our purpose here, then, is to re-examine some of the data and present material not covered by Shuey, rather than to repeat what Shuey has done with a fair degree of thoroughness.

McGurk's studies. Evidencing a long time interest (McGurk, 1943) in Negro-white comparisons of intelligence, McGurk (1951, 1953a, 1953b) has received considerable publicity in both scholarly and lay circles because of research which seems to undermine the thesis that cultural factors rather than native endowment factors account for differences between whites and Negroes in intellectual functioning. Two hundred thirteen Negro and 213 white youths were matched for age and curriculum and for social and economic variables by means of a shortened form of Sims' Record Card. McGurk found that even when socioeconomic variables were thus controlled, the mean scores of whites and Negroes differed significantly in the direction usually reported. Also, when the specially devised test items were separated into "cultural" and "noncultural," the differences were greater on the noncultural questions than on the cultural questions, contrary to environmentalist indications. Further, as socioeconomic status increased, the differences between Negroes and whites increased rather than decreased, again contrary to expectations from a cultural theory of differences in intelligence.

One point at which McGurk's analysis might be misleading to the statistically untrained has been pointed out by Long (1957). McGurk states that 25% of Negroes overlap whites, from which it might be concluded that only 25% of the specified Negro population have scores in common with the specified white population. Actually, analysis of McGurk's data shows that 91% of the Negroes have scores in common with the whites. Inasmuch as McGurk has addressed himself to the lay public, he should make abundantly clear that "overlap" is used in the technical sense of exceeding (or for an upper distribution, falling below) the mean or median of another group.

McGurk does not claim, though he seems to imply, that the superiority of whites on intelligence tests results from innate factors. In the final analysis the implication may be correct, but research such as McGurk's cannot establish it. Indeed, we do not see how the issue can be resolved by any number of ingenious methods of equating for social and economic variables. The various indices of socioeconomic status already devised or those at present conceivable on the same principles are intended to distinguish social *classes* from one another. How they can be employed to compare individuals in different *castes*, except very roughly, is difficult to see.

In actuality, not only in the South (Dollard, 1949), but North as well

(Brown, 1944; Long, 1957) whites and Negroes comprise separate castes; they are not merely representatives of different classes. In the state of Florida where the writers reside there are a number of Negroes whose social and economic statuses exceed those of most white persons. These Negroes, however, cannot yet sit in the same seats on public transportation (in most places), go to the same hotel, restaurant, club, school, church, social events, or even restrooms. Although some of these strictures do not hold in Northern states, attitudes regarding intermarriage and the more personal forms of social intercourse do not appear greatly different from those held in the South. From the Early Childhood Project (Radke, Trager, & Davis, 1949) we learn that in Pennsylvania, from which state McGurk drew part of his sample, children discern within at least the first four or five years of life their social and ethnic roles, with attendant supervaluations or devaluations of self and performance expectations. Interlocked with caste variables are those which influence performance, such as the color of the investigator (Trent, 1954), which in turn may be related to deterioration of intellectual performance under anxiety-provoking conditions (Beier, 1951; Hammer, 1954).

We wish to emphasize that we are not taking sides at this point in the heredity-environment controversy in relation to intellectual differences; we believe both camps (e.g., Canady, 1943, for environmentalists) have mistakenly assumed that if the two racial groups are equated in terms of social class and economic variables that a definitive answer can be found concerning even the *relative* weighting of innate or acquired factors. Involved here are different dimensions (possibly correlated, of course) not merely different quantities along the same scale. Quibbling with McGurk over minor points of methodology should not obscure the value of the valiant attempt he has made to test hypotheses proposed by the environmentalists. The error lies in the assumptions both he and his opponents make.

Studies on the WISC. Only one investigation (Young & Bright, 1954) is cited by Shuey under WISC studies. Not surprisingly in view of its standardization, the WISC was found inappropriate for testing Southern Negro children from 10 to 13 years of age.

Another more extensive study has come to the writers' attention, that of Caldwell (1954). Four hundred and twenty Negro children were tested ranging from 6 to 12 years of age, with equal numbers of males and females, drawn from towns in five deep South states and randomly selected from school rosters. One examiner tested 342 of the subjects. According to the report, "excellent rapport was obtained," although the means of establishing rapport appear inadequate especially in the light of the conspicuous Southern accent of the chief white examiner. Caldwell's hypothesis that a difference exists between Southern Negro children and the white standardizing group was, as might be expected, borne out. Socioeconomic class influence was probably strong, with 75% of the subjects in the lowest third of SE groupings. Nevertheless, the Full Scale IQ mean of 85.52 is considerably higher than that obtained in the Young and Bright study (mean = 67.74). (Something is wrong with the standard deviations reported on p. 18 of Caldwell's dissertation: *SD*s ranging from 0.63 to 1.4 sound more like standard errors, but do not seem to jibe with either the WISC Manual's, the data

of Young and Bright (1954), or the spread one might expect.) In this investigation the suggestion is also made that cultural bias results from using the WISC, standardized as it was on a white population.

"Culture-free" or "Culture-fair" studies. McGurk's attempt to differentiate cultural and noncultural tests from one another is only one in a series which began with the use of the Army Beta in the first World War. Although the first performance tests were intended to overcome literacy handicaps, they became in time useful for estimating intelligence across ethnic, racial, and cultural barriers. In Table 2 are summarized endeavors to compare Negro and white children by means of tests which were either, like the Beta, adapted for cross-barrier use, or specifically constructed to eliminate cultural factors. Some of these studies are cited by Shuey, but are included here as essential for understanding this area of comparison.

As far as we have been able to tell, no one has employed Cattell's "culture-free" test (Cattell, 1940, 1951; Cattell, Feingold, & Sarason, 1941) for comparing Negroes and whites. This test has yet to prove itself, however, as even "culture-fair," to say nothing of "culture-free." Marquart and Bailey (1955) present convincing evidence that the "culture-free" test has sufficient variance associated with the Binet to suggest they are measuring much the same capacities. Scale 1 correlates with the Binet .74 and Scale 2 correlates .81. Most test constructors who would want to measure what the Binet measures would be very happy to obtain such coefficients. What Goodenough and Harris (1950) say about the Draw-a-Person Test appears relevant to all "culture-free" or "culture-fair" tests: The search for a culture-free test is illusory.

TABLE 2

COMPARISONS ON "CULTURE-FREE" OR "CULTURE-FAIR" TESTS

Author(s)	Test	Subjects	Results	Comments
Woods, Boger, & Holman (1954)	Beta (Lindner-Gurvitz Scale)	N adolescents, delinquents, nondelinquents	Subtest 4 low	Authors conclude aim of subtest equality not fulfilled or cultural factors deprive N on subtest 4
Woods & Toal (1957)	Revised Beta	N & W adolescents matched for IQ	W > N on Paper Formboard, Drawing Completion; N ≥ W on Digit Symbol, Visual Comparisons	Authors suggest Detection of Errors and Drawing Completion are culturally loaded—probably *ex post facto* theorizing
Love & Beach (1957)	Davis-Eells Games	N & W from three S-E levels	W > N	Differences obscured by N's being in two lower S-E levels only
Fowler (1957)	3 "culture-controlled" tests	N & W, ethnic & nonethnic	W > N	Positive correlation also found between conventional and "culture-controlled" tests and S-E status
Hammer (1954)	CTMM (nonlanguage)	N children and adolescents	N < W norms	Results same with both language and nonlanguage factors
Newland & Lawrence (1953)	Chicago Non-Verbal Exam.	E. Tenn. N. School children	N < W norms	Differences 2, 3, and more years
Coppinger & Ammons (1952)	Full-Range Picture Vocabulary Test	N children from 2 La. Parishes	N < W about two years	At upper age levels urban N children's mean scores ≥ W norms.
Hess (1955)	"Culture-fair" test of 16 items	High & low status W, low status N	W high status > N	No difference between W groups. Smaller difference between N & W on "culture-fair" test than on standard intelligence test

Gifted children. Shuey's review of comparative studies (both directly and by implication) reveals that whites produce a greater proportion of gifted children by far than do Negroes. The percentage reported by Shuey for the latter, about 0.14% or 0.15% is well below the 0.95% for white children testing 140 IQ or above on the Binet or comparable scales. This low proportion among Negroes is only an expression of the general situation, i.e., that the whole curve for Negroes on most intelligence tests is displaced downward. Even taking Jenkins' (1950) figure of 0.3% for the 140 IQ or above places the area under the upper end of the curve below that for whites.

In surveying cases of gifted Negro children, Jenkins at first (1943) maintained that the gifted Negro child has essentially the same characteristics as a comparable white child. But later Jenkins (1950) recognized from more intensive investigation that the most important single fact for any Negro, gifted or not, is his being a Negro. Consequently, the performance he manifests on an intelligence test as well as elsewhere is literally colored by this fact.

That Negro children earn IQs of 160 or above (Jenkins, 1943) or even as high as 200 (Theman & Witty, 1943; Witty & Jenkins, 1935) on tests standardized on whites is a remarkable phenomenon. Anyone who has tested Negro children has probably been impressed with the fact that a Negro child whom the examiner knows to be functioning as a normal, not a retarded child, may receive a score which automatically would classify him in the retarded range if scores only were regarded. Roughly speaking, the Negro child seems to operate in everyday life situations in a way expected of a white child about 10 IQ points above. With the curve of measured Negro intelligence displaced downward it is thus a surprise to find any Negro children scoring among the highest levels on white-standardized tests.

Specific abilities. There have been few studies comparing Negroes and whites on specific abilities. Only two such reports worthy of mention came to the attention of the writers. Prothro and Perry (1950) administered the Meier Art Test to 460 high school and college students in Louisiana and found that mean score comparisons placed the whites above the Negroes. However, the authors felt that the differences might be due to the socioeconomic status of the groups. Negro and white sixth graders were tested on the Kwalwasser-Dykema Music Ability Tests with the Negroes obtaining a higher median score (Woods & Martin, 1943). The authors conclude that the Negroes were superior in a number of specific areas, but the meager statistical data presented do not support their conclusions.

Klugman (1944) in a study not mentioned by Shuey discovered that money incentives and praise had no demonstrable effects on white children (CA 7–14) in taking the alternate form of the Binet after an initial first form. However, comparable age Negro children given money rewards showed better performance than those given praise as an incentive. (Incidentally the two groups averaged 99.11 for whites and 98.00 for Negroes, with 6.08 and 5.66 *SD*s, respectively.)

In an investigation which compares with similar studies on white children, Robinson and Meenes (1947) found the relation between Negro children's test intelligence and parental occupation very low. From 1938–39 to 1945–46, the intelligence average rose from 97.0 to 99.7, a

statistically significant change, and the relation between occupation of parents and intelligence increased as well—in this case it might be better to say that the relative lack of relation became less pronounced. A greater number of Negroes in higher status occupations could increase the coefficient attenuated by abbreviated spread of scores.

Davis (1948) and Davis and Havighurst (1946) in a test of the influence of social class upon learning draw attention to differences in child-rearing practices of lower and middle classes, in general, that less rigid behavior is expected of children in lower classes. Comparing 100 Negro and 100 white families, Davis first indicates that the Negroes tended to be in the lower class. Where Negro families could be regarded as middle class, their child-training practices were conservative in regard to feeding, toilet training, and masturbation. It seemed to Davis that differences in adult behavior were greater within either Negro or white groups than between them. On mental tests, in which our interest lies principally, there were significant differences between the high and low socioeconomic groups. Davis seems to assume that these differences are culturally determined, but gives no satisfactory evidence in this regard. He does go on to urge the development of "culture-fair" tests and the elimination of academic intellectual tasks in the testing of intelligence.

Many years ago Sunne (1917) discovered that on some items of a test Negro children excelled whites and on others whites excelled Negroes. Sunne gave no satisfactory explanation for the finding. We have cited several studies above which have produced the same results. At least part of the differential performance may be attributable to status differences if the lead suggested by Eells, Davis, Havighurst, Herrick, and Tyler (1951) is in the right direction. These investigators found that status differences among children affect some test items and not others.

Adults

In addition to Shuey's review (Shuey, 1958), we report here a few other studies from the military services and other special groups for which comparisons have been made. One massive investigation, part of the standardization of the PAT (Tomkins & Miner, 1957), is not mentioned by Shuey, but by all means should be brought to the reader's attention because of the careful selection of representative subjects. The investigator (Miner, 1957) employed a 20-word vocabulary test adapted from the CAVD. The sample of 1500 individuals was chosen from stratified random clusters of blocks or rural areas with quotas on important variables. Whites averaged 11.06, *SD* 3.41, and Negroes 8.08, *SD* 2.72, a significant difference.

Armed Forces

In the period since 1944 debate over testing of intelligence in both world wars has been carried on. Although it seems to have subsided, it is not so much that the issues have been settled as that the contestants have become exhausted. To Shuey's conclusion, that in both wars whites consistently did better than Negroes on the average, few would take exception. She takes up one by one the reasons advanced by others for the discrepancies and provides reasons on the opposite side which purportedly answer an egalitarian view. The reader will have to study Shuey in order to judge whether the data support her reasoning. Additional dis-

cussion of some of the issues involved can be found in a series of articles by Garrett (1945a, 1945b, 1945c, 1947).

Several studies not cited by Shuey are mentioned here (Altus, 1946; Altus & Bell, 1947; Fulk & Harrell, 1952; MacPhee, Wright, & Cummings, 1947), primarily for the sake of completeness inasmuch as they do not differ in their results from others. Fulk and Harrell endeavored to equate groups by the last school grade completed; they compared AGCT scores and discovered that whites were favored at every school grade completed. The authors recognized that equal schooling does not render groups equivalent. Origins of their subjects, a highly important variable, were not controlled. Altus and Bell interpreted their discrepant findings on illiterates as due to cultural factors.

In a survey of "the uneducated" Ginsberg and Bray (1953) cite figures to indicate that the number of Negro illiterates declined from three million to slightly more than one million from 1890 to 1940. In World War II, 391,300 whites and 325,100 Negroes were rejected on the grounds of illiteracy, almost three-fourths of these rejectees being from the Southeast and Southwest.

Ginsberg (1956) draws the conclusion from studies of manpower in the Second World War that there is a great wastage of human potential when one realizes that the absolute numbers of Negroes in the upper classes of the General Classification Test are large, even though percentagewise Negroes have small representation in Classes I and II. The late Walter V. Bingham pointed out the waste of potential of all groups implied by the fact that one-fourth of the truck drivers exceeded one-fourth of the bank executives.

Special Civilian Groups

A large number of comparisons of college students reviewed by Shuey can be supplemented by only one series of studies by Roberts (1946, 1948, 1950). Northern Negro college students did better on the ACE than did their Southern counterparts. In a longitudinal comparison original difference in academic achievement total scores were erased after four years of college, even though ACE scores still reflected differences. During the four-year period the subjects made greater gains than those expected by the national norms for a similar period of time.

Comparing Southern Negro and white venereal disease patients, Scarborough (1956) discovered differences on the Wechsler-Bellevue paralleling those of other investigators, but the differences between white and Negro VD patients did not seem to be as great as those between white and Negro controls. Davis (1957) using the same test found no difference between 33 mental patients and 27 controls, though the overall IQs of 67 and 68 are well below white norms. Also employing the Wechsler, DeStephens (1953) answered his own question, "Are criminals morons?" in the negative when he discovered that 200 white and 100 Negro admissions to an Ohio reformatory yielded the following average scores: FS: W 93.55, N 87.90; V: W 90.13, N 86.70; P: W 98.30, N 91.20. On the basis of the standard deviations reported none of these differences between groups is significant. Findings of Reitzes (1958) on Negro applicants to medical schools yield a different result. Generally it is expected that Negro applicants will rank lower than white applicants on the Medical Colleges Admissions Test. Regional differences appeared among Negro applicants with North and West

above border applicants, who in turn ranked higher than Southerners.

Overview of Intellectual Functions

Several grave issues arise in a review of comparisons of intelligence of whites and Negroes. We shall phrase these and make a few comments on each. It will be obvious that some of the problems apply more generally than to intelligence comparisons, but they are set forth here inasmuch as all of them are related to the foregoing comparisons.

What Constitutes a Race?

For convenience we have assumed that groups belong to the white "race" or Negro "race" as designated by the investigator. Any research, however, which seriously attempts to make comparisons between or among "races" must sooner or later grapple with the problem of this section. Skin color, hair texture, and other physical characteristics have proven illusive as definitive criteria. We cannot settle a question to which anthropologists appear not yet to have given an adequate answer. Nevertheless, psychologists of the future must make explicit the concept of race to which they adhere in order to make clear what populations they are comparing.

Determination of Racial Composition of Groups

Assuming some acceptable definition of race, the investigator needs to determine in each case whether the groups he is comparing actually are differentiated by his criteria. If, for example, he has adopted heredity, the possession of such-and-such a percentage of accepted white or Negro ancestry, as a criterion, the researcher must make certain his groups fall within the limits of the specific heredity he has accepted. One suspects that in a number of cases so-called racial comparisons are being carried out between one group designated as "white" and another designated as "black" which consists of many who are partly or even largely white.

Confusion of Class and Caste Variables

As we have indicated in commenting on McGurk's studies, both hereditarians and environmentalists have fallen into the trap of assuming that if they can get two groups who are equated in socioeconomic terms there can be an *experimentum crucis* which answers the question of what parts heredity and environment play in intelligence differences. Canady (1943a) discussed the great difficulty in equating environments, and Anastasi and Foley (1949) pointed out that formal education and socioeconomic differences were not enough to account for differences in tested abilities of northern and southern Negroes. Actually, we despair of being able to equate groups until caste differences are removed and only class differences remain.

Interaction of Examiner and Subject Variables

In some studies (Trent, 1954) the color of the examiner has been taken into account in assessing results; in others this factor has been ignored. Repeated testing of the same individual, both white and Negro, by both white and Negro examiners would seem to be called for in order to determine the proportion of variance in intelligence test scores attributable to interaction of examiner and subject.

The Functional Value of Intelligence

What kinds of intelligence do white persons need and what do Negro persons need to survive, adjust, and

make progress? It is naive to assume that the academic types of intelligence tests which have traditionally been the instruments of comparison compare in reality Negroes and whites in those areas of intelligence which they are called upon to use in "real life" situations. Intelligence test differences between Negroes and whites cannot mean the same as they mean between two groups of whites.

If we assume that intellectual functions develop adaptively and are not entirely determined by heredity, we may suppose that intelligence tests of the usual variety measure in part that which is developed in order to achieve success in a certain culture. A negro in a white man's world requires a kind of intelligence enabling him to detect from minimal cues how a white man is going to react to a critical situation. Usually his success and sometimes his life depend upon this kind of intelligence, hardly ever upon whether he can define "ethnology." Tests of intelligence tapping the kinds of intellectual functioning called for in achieving success in the actual world men have to face might reveal different results in comparative studies, not only between Negroes and whites, but other groups as well.

Newer Concepts of Intelligence

Practically no research has been done comparing white and Negro subjects on factors of intellect (Guilford, 1956, 1959; Thurstone, 1938). The only studies of this kind we have found are those by Lee (1951) and Michael (1947). Even studies with performance-type tests, which in a sense go outside of academic intelligence quotient concepts, do not meet the need to reach the many factors revealed in modern factor-analytic research. A world of research is called for to determine how the two groups we are considering compare on the 50 or more factors of intellect. Caution needs to be exercised here in respect to the functional value of even Guilford's many factors. These were derived from tasks set for high-level personnel in the military and might not have anything specifically to do with what a person in a different class and caste needs to do. However, *insofar* as these factors cover the entire range of intellect, comparisons of factors should be much more valid than of gross IQs.

Significance of Overlapping Distributions

A legitimate question arises in statistics in relation to assigning individuals to one of two (or more) distributions (Horst, 1956). In statistical theory individual scores in a distribution may be regarded as errors departing from a mean (Yule & Kendall, 1949). When two distributions are compared, then, it is assumed that individual scores are only errors of observation from their respective means, which in turn are regarded as estimates of either one true population mean (if the null hypothesis is not rejected) or of two true population means (if the null hypothesis is rejected). In the case of intelligence comparisons the null hypothesis may be stated: the obtained means of these two distributions of intelligence scores differ only by chance, i.e., they are really only estimates of the same true intelligence score population mean. On the whole, investigators have rejected the null hypothesis in comparing Negroes and whites and have concluded that the two means represent two different intelligence population means.

In connection with any one individual, however, especially in the area beyond the mean of either distribution, the question may legiti-

mately be asked: To which *intelligence* mean does he belong—or, from which *intelligence* mean is his score a deviate? It seems to be assumed that because an individual is white or black, his score must therefore be a deviate from a *white* intelligence population mean or a *black* intelligence population mean. But this assumption begs the question. The individual may be a deviate from the *lower* intelligence population mean whether he is white or black, or from the *upper* intelligence population mean whatever his color. The individual who is white but in the very low part of the scale may in truth be a deviant from the upper mean but we cannot know this fact merely because he is white. We could only be relatively sure of this fact if there were no overlap in the absolute sense between the two distributions.

This statistical consideration is a variant of the more general one which suggests that other factors than color (i.e., genetically determined intellectual concomitants of color) decide whether an individual makes a high or low score on an intelligence scale. A statement of the statistical situation focuses attention on the overlap found in virtually every study and what it must signify in relation to any one individual.

Social Consequences of Research Findings

In the introduction to Shuey's book (1958), Garrett discusses the need not to interpret Shuey's conclusion as a basis for differential treatment of the two groups, especially for mistreatment of Negroes. A view strictly limited to scientific conclusions would be, "Here are the results of examination of the data. It is not our responsibility to recommend courses of action." If there should be significant differences between groups which can be shown to arise principally from genetic factors, the practical response of the man on the street (white or Negro) would almost inevitably be to justify either his treating others as inferiors or to accept his own position of inferiority as natively determined. We are not convinced that genetic differences have been shown; but even if they were so shown, we believe it is incumbent upon the social scientist to set forth the full picture. The wide overlap between white and Negro distributions of scores should be pointed out so that it is evident that within group differences are far greater than between group differences. It should also be shown that oftimes two groups of white persons differ significantly, and probably in some if not all cases, partly because of genetic factors. Social scientists need to be alert to the implications of their findings.

Educational and Post-Educational Attainment

Table 3 presents mostly direct Negro-white comparisons in school achievement and in the case of Rodgers' study some post-educational attainment. It is generally recognized that as a national group Negroes receive a poorer education than whites both in quantity and quality, a condition reflected in scores on achievement tests. Other factors, however, may have to be adduced for explaining Witty and Theman's results.

Temperament[3]

The studies in the following paragraphs are subdivided according to

[3] Some would use "personality" rather than "temperament." We follow what seems to be the trend in recent years, to regard personality as the larger set of functions under which temperament is subsumed. The latter

TABLE 3

RELATIVE ATTAINMENT OF NEGROES AND WHITES[a]

Author(s)	Subjects	Results	Comments
Thompson (1936)	College teachers	Pay, morale: W > N teaching load: N > W	Author points out little research done by N
Ferrell (1949)	Grades 4, 5, 6	W > N on Stanford Achievement Test (all areas)	W more variable
Bullock (1950)	High school graduates	W > N on Iowa High School Content Exam.	Some question about choice of samples but same results regardless of type of analysis
Witty & Theman (1943)	N youth, Binet IQs 120–200 identified 6 years before	W gifted < N > average	Authors state attainment other than educational relatively high
Rodgers (1957)	Baltimore superior and average N identified in 1933	Gifted superior in achieving middle class, education, etc.	Comparison with W only indirect

[a] Bradley (1949) discusses relative literacy rates among military selectees, and Davenport (1946) educational attainment as it relates to military selection.

the instruments employed. We sought for a more logical set of divisions, but fell back on this artificial, yet serviceable classification. One major reason for employing the test as a separate category is that the reader may judge the validity of the studies partly in terms of his knowledge of the validity of the test. With the exception of the Murray TAT, almost all of the instruments utilized for comparative research have been gravely questioned by qualified sceptics. If studies using some particular test are scattered throughout a section among other results, it would be more difficult for a reader to judge whether the results from the particular test reflect the validity or invalidity of the test or stem from the experimental conditions. This consideration is not a major one in other parts of this review.

Rorschach

It seems to us that there must be more studies comparing white and Negro on the Rorschach, but we have found only two. Morons from both groups were compared by Abel, Piotrowski, and Stone (1944) in terms of specific responses to the ink blots. Out of the entire set of comparisons only one showed a real difference between the two groups: Negroes gave more *M* than whites. What the meaning of this finding is depends first upon whether this is a chance result in comparison with the probabilities involved, and/or secondly, if it is not chance, upon the meaning of *M*.

The group Rorschach was administered by Stainbrook and Siegel (1944) to high school and college students. No sampling statistics are presented, but responses are handled on the "simple basis of the probable differences of mean frequency of occurrences." Various determinants were studied in isolation. The authors conclude that both high school and college Negroes show less fluidity in association, that high school Negro youth are more emotionally stable and less impulsive than their white counterparts, and possess less anxiety, but that white college students though showing more emotional irritability are more mature, possess more "general personality resources"

includes traits or reaction systems like cyclothymia-schizothymia, introversion-extraversion, neuroticism, and others which are largely of constitutional origin and not so much affected by environmental changes as other functions of the personality such as attitudes and social interests.

than Negro college students and have a "more daring intellect." Inasmuch as the group Rorschach has better norms but less interpretability than the individual Rorschach, it is difficult to know how to evaluate this investigation in comparison with most Rorschach studies which use the individual form. Harrower-Erickson and Steiner (1945) warn against using interpretations from individual Rorschach practice for group Rorschachs.

Thematic Apperception Test

Murray TAT

With a white female examiner Abel (1945) found that both white males and females and Negro females were more communicative on the TAT than Negro males, who were of at least equal intelligence to the other groups. Communicativeness was measured by the number of ideas and the number of words, both of which yielded significant differences. It scarcely seems surprising in the light of other research on the TAT and some knowledge of Negro-white sociodynamics that Negro males were inhibited in the presence of a white female examiner; but possibly we are engaging in ad hoc reasoning. At any rate, the author failed to point out a major disclosure of her data, that sex differences were greater than racial differences on the measured variables.

Differences between TAT responses of Negro and white boys were investigated by Mussen (1953). Fifty of each group from lower-class, New York City homes, of at least average intelligence, and between the ages of nine and 14 were age-matched. They were given 13 TAT cards plus one special mother-child card and their responses were analyzed by a modified Stein Schema with 28 need and 22 press categories. Scoring was blind. On most of the items the two groups were similar, a not unimportant finding. Nevertheless, 14 out of 50 chi squares were significant at least at the 5% level of confidence, a more than chance number of significance tests. Negro boys expressed greater hostility in thought processes than did white boys, less need for acting out murderous aggression, but about the same for other types of acting out. The Negroes manifested more need to reflect, think and speculate, but showed less desire for establishing and maintaining friendly relations, expressing admiration and respect for others, or being respected, followed and obeyed. They displayed less n Ach. Although white boys regarded others as being more rejectant than Negro boys, they also looked upon them as being more friendly, while the Negro boys viewed the environment as more hostile. On the whole, it appears from Mussen's data that the self-attitudes of the Negro boys and their corresponding attitudes towards others have suffered somewhere along the line, even though in most things their temperaments parallel those of white boys.

More research utilizing the (Murray) TAT is called for to make comparisons of temperament variables. Of all the instruments employed for extra-intelligence investigation herein mentioned the TAT has suffered least at the hands of critics. It has demonstrated its value both clinically and experimentally. Consequently, comparative findings are not subject to the double jeopardy of having both methodology and test instrument attacked. The TAT has its faults in terms of interscorer reliability, but it is a fairly well proven technique for uncovering psychodynamics.

Thompson TAT

In a logically-motivated and commendable effort to make it easier for Negroes to identify with TAT characters, Thompson (1949) published his Negro version (T-TAT) with characters of obviously Negroid features. For some reason Thompson's ill-starred attempt has stirred more research than one might expect, all of which appears to invalidate the assumption that Negroes identify better with pictured characters of their own race than with corresponding white characters. Incidental to the attempts to refute the assumption are direct or implied comparisons of Negroes and whites on the T-TAT or M-TAT.

Korchin, Mitchell, and Meltzoff (1950), utilizing both T-TAT and M-TAT, performed an analysis of variance with race and social class as independent variables. No significant differences were found between two Philadelphia racial samples, but there were significant differences between social classes on length of Murray TAT stories. Korchin et al. believe that the assumptions underlying the T-TAT are not justified, so that a Negro modification is not called for. In a fairly well-designed experiment Riess, Schwartz, and Cottingham (1950) reached the same conclusion concerning the Thompson assumptions and strengthened their case in another study (Schwartz, Riess, & Cottingham, 1951).

Four groups of subjects, two white and two Negro, were given the M-TAT and the T-TAT by Cook (1953), the former to one group from each race and the latter to the alternate group from each race. Confirming previous findings on the lack of necessity for a Negro modification of the TAT, Negroes regarded characters in both sets of pictures as representative of people in general. Whites, on the other hand, looked upon the Thompson characters as Negroes rather than as persons-in-general. As a whole, considering the production on both sets, Negroes offered a significantly larger number of alternatives and of words indicating uncertainty, a higher vagueness score, and a large number of references to the pictures as pictures. Whites gave a higher word count.

Carrying out research on whites only, Light (1955) found no significant differences between his matched subjects in the productions on the M-TAT and the T-TAT presented in balanced order. Light's subjects, like Cook's, responded to the T-TAT characters as Negroes, with 14 of 26 subjects using traditional themes of Negro inferiority. One difficulty in interpreting these results comes from our not being able to know whether the examiner was white or Negro.

Although Thompson's attempt to provide easier identification for Negro subjects on a projective test has not borne its intended fruit, the result of comparing whites and Negroes seems to be a highlighting of the cultural pattern of virtually universal use of white characters as illustrations, at least in publications white people see. Negroes perceive white and black as people; whites tend to see white as people and black as Negroes. Whether the explanation is as simple as that of experience with the cultural pattern remains for future research to disclose.

Picture Arrangement Test

The Tomkins-Horn Picture Arrangement Test developed by Tomkins and Miner (1957) is probably the most extensively standardized projective test in the United States

if not in the world. We are not prepared to say that it is the most adequately standardized, for we are not certain Gallup Poll interviewers are qualified to administer even an objectively-scored projective technique. However, the more than 1500 persons on whom the norms are based were chosen by block clusters and quotas in approved opinion sampling style. In addition to administering the 25 plates of the PAT according to rigidly standard procedures, the interviewer obtained an intelligence indication by means of a 20-word Thorndike-Gallup vocabulary test, together with information on major demographic variables.

The test itself consists of a series of plates with three positions of one or more persons arranged in a circular fashion so that the examinee must either stand on his head or turn the plate around in order to get a right-side-up view of all three positions. Scoring is objective—it may be done by a clerk—and is intended to reveal information on 158 temperament variables. Rarity of responses, determined empirically from the normative sample, yields an individual's "variant" score which in turn varies depending upon his age, education, and vocabulary IQ.

With the PAT which he helped to standardize in part, Karon (1958) performed an extensive series of analyses, utilizing a special discriminant function, to compare Negroes and whites with the intent of finding what effect, if any, caste sanctions have on Negroes. Karon's book is very confusing in its description of procedure. As far as we can ascertain, three "validation" samples of Northern and Southern Negroes and Northern whites were selected partly from among the 1500 persons mentioned above and compared. Final samples were chosen from ninth grade Negro students from North and South.

Results of the analyses confirmed in each case, according to Karon, the hypothesis that caste sanctions affect adversely the "personality characteristics" of Negroes. Significant differences between discriminant score means are interpreted as the result of differential caste sanctions. These differences appear to indicate that actual possession of and simultaneous denial of aggression is the principal problem of the individual living under caste restrictions, with a secondary resultant of flattened affect. There are differential "human costs" even between areas of more severe and those of less severe sanctions.

Some methodological umbrage could be taken in view of Karon's use of subjects taken from the standardization group for part of his experiment. Nevertheless, the results of his pilot studies were validated on what seem to be entirely independent samples.

Another possibility exists for interpreting Karon's results. Some form of selective migration explanation could apply to the dispersal of Negroes which could account for differences in affectivity and reactions to aggression. It is possible that those Negroes who have remained under the severest caste sanctions have remained because of their temperament characteristics and those who have stayed under less severe restrictions or moved to relatively caste-free environments have done so because of their characteristics. In other words, they (and their children who presumably resemble them) are under or out of caste sanctions on account of their personalities rather than having certain personality characteristics on account of sanctions. Those who can swallow their aggressions—with occasional outbreaks against their fellows—stay, those who cannot, go. This explanation seems fully as plausible as the

other, although we recognize the two are not mutually exclusive.

Picture-Frustration Study

Rosenzweig's Picture Frustration Study has been employed in several comparative investigations which are summarized in Table 4. There seems to be a tendency for Negroes to project more aggressive responses (E) than do whites. On the whole, similarities of responses are much more pronounced than differences, a situation which may merely reflect the relative insensitivity of the instrument.

Miscellaneous Tests and Other Instruments

In 1942, on the basis of intercorrelations of several scales administered to Fisk University students, Negro college students were regarded by Brunschwig (1942) as appropriate subjects with whom to employ tests and ratings devised for whites. Whether scales standardized on one group can be applied with equal validity to another racial or ethnic group is still a question. The TAT studies cited above might suggest that they can be. Whether or not they can be, we report studies from uncritical applications and also from unstandardized instruments. Roughly, this portion of the review progresses from children to adults and chronologically.

Reporting results of the same research in two journals, Hammer (1953a, 1953b) compared children in grades one through eight on the H-T-P Test; no control was exerted over socioeconomic variables. Clinician judges ranked the drawings on a six-point scale of neuroticism and a three-point scale of aggression. Negro children rated higher (at the .01 level) in aggressiveness. And whereas white children ranked between mildly neurotic and neurotic, Negro children scored on the average above severely neurotic. At every grade level, Negroes ranked higher in neuroticism, although at the upper levels white children became relatively more neurotic, thus decreasing the difference. Negroes showed specifically more

TABLE 4

COMPARISONS MADE WITH ORIGINAL AND MODIFIED ROSENZWEIG PICTURE-FRUSTRATION STUDY

Author(s)	Test	Subjects	Results	Comments
McCary (1951)	PFS	Northern and Southern N and W	SN women > NW women on E[a] compared with I[a]	
McCary (1956)	PFS	Northern and Southern N and W, 14–22 yrs.	Most scores as; NW and NN > SW and SN on E compared with M[a] and I	Normative study—samples should not be used for experiment
McCary & Tracktir (1957)	PFS	High, medium, and low IQ W and N	Low IQ: N > W males on E compared with I and GCR[a]	
			Medium IQ: W > N females on N-P[a]	
			High IQ: N > W females on O-D[a]	
Portnoy & Stacey (1954)	Children's PFS	Subnormal, matched N and W children	N > W on M	Matching: length of institutionalization, intelligence, etc.
Winslow & Brainerd (1950)	Modified PFS (W or N frustrating agent)	Matched N and W	Both N & W: E more with N frustrator	Matching: age, sex, education, S-E[b] status

[a] E = Extrapunitive, I = Intropunitive, M = Impunitive, GCR = Group Conformity Rating, N-P = Need-persistence, O-D = Obstacle-dominance.
[b] S-E = Socioeconomic.

bleak trees, more concern with the house, and more organic signs.

Disregarding the lack of established validity for the H-T-P, we consider the clinician's ratings as too severe. Possibly, as Willoughby observed many years ago, what we call normality may only be a widespread case of arrested development; nevertheless, the students on the whole in Hammer's study were probably not neurotic according to ordinary standards. Should the H-T-P be valid, an investigation controlling for class and caste variables is called for rather than the type reviewed here.

Among Letchworth Village mental defectives, it was found (Abel, 1943) that white and Negro girls, matched on relative level of IQ, more or less leadership ability, and other relevant variables, differed in that the Negro girls displayed more dominant behavior in imposing their judgments upon the white girls and making the decisions. Boyd's (1952) results, working with children of more or less normal intelligence in a nonsegregated elementary school, may be juxtaposed with the above. On two tests and a questionnaire designed to determine levels of aspiration, matched groups (age, IQ, economic status) revealed on the target test, questionnaire, and on the arithmetic test, higher Negro 1-o-a. In selecting "the greatest person in the world" and "the person I would like to be like," 24 of 28 persons chosen by the Negro children were Negroes.

The above two studies are placed together for the possible implications of compensatory behavior elicited in a society dominated by one group to the real or apparent detriment of the dominated group. Of course, the institutional behavior may not be compensatory acting out, but could be explained at least in part on the basis that the Negro girls are not as defective in reality as their white IQ counterparts (cf. the discussion elsewhere in this article on actual versus measured intelligence of low IQ Negroes).

Gray (1944b) made a study of the wishes of Negro elementary school children and compared them with those of white children from a previous study. The most striking point is the basic similarities of the two groups, although Negro children expressed more wishes concerned with home, animals, and musical instruments. Doll play fantasies of Negro and white children may be inferred comparatively from an interracial comparison by Graham (1955). Individuals ranged from 73–102 months and were regarded as homogeneous within groups. Although the results are difficult to evaluate interracially, it appears that the 30 Negroes produced fewer total fantasies than the 30 white children. With some exceptions, both groups produced approximately the same proportions of stereotype (dining room, kitchen, etc.) and nonstereotype (affection, aggression, etc.) responses.

One other study of children gives by implication some meager comparative information from a temperament test. Anderson (1947) administered two group intelligence scales the Otis and California Test of Mental Maturity, and a group "personality" test, the California Test of Personality, to 153 Negro pupils in an Oklahoma high school. He found the averages on the intelligence scales at or above the means expected, but the Negro youth low in personal and social maturity on the CTP. The author does not mention the question of standardization of the CTP on a white population. The intelligence test results can be cautiously interpreted as running counter to the general findings elsewhere.

Illiterate soldiers revealed sectional differences on the Altus Adjustment Scale according to Altus and Clark (1949). With a higher mean signifying poorer adjustment, the following means were obtained in this study:

Southern Negro	7.92
Northern white	9.40
Northern Negro	9.56
Southern white	11.26

An analysis of variance suggests that these differences are not the result of chance factors.

Contrary to Altus' findings with his scale, Felton (1949), employing the Cornell Selectee Index with 148 Negroes and approximately 2400 whites at Oak Ridge, discovered that the former were significantly more frequently in the neurotic group. The author does not break down the figures, but a direct relation shows up in the data between lack of education and neuroticism. Although the Negro group is described as not from the deep South, there is a possibility that they were in the less well-educated groups. Touchstone (1957) found that on a sentence completion test neither white nor Negro rated higher than the other in passivity, aggression, hostility, or withdrawal, on all of which the investigator had expected to find Negroes scoring higher.

Overview of Temperament Studies

Living in a white culture the Negro seems to have difficulty with his frustration-induced aggression, although some reversals (which may to be sure, be test artefacts) suggest that even under severe caste restrictions aggressive drives may not be the most important problem in the Negro's handling interracial interactions. It seems to be a truism that more and better research is called for in any area of investigation, but especially here where norms have scarcely begun to be established, where tests which have even failed to prove themselves with white subjects are applied almost unquestioningly to Negro subjects, and tests with white norms are utilized as if class and caste distinctions have no bearing on temperament or personality as a whole.

So-called "personality tests" may be inappropriate for testing most Negroes who are different from whites in socioeconomic status (Auld, 1952; Hoffman & Albizu-Miranda, 1955) and caste. In this area intensive studies of whites and Negroes need to be performed by scientists who understand both psychodynamics, sociodynamics, and adequate scientific procedure.

Values and Attitudes

Various attempts have been made to measure the value systems and attitudes of groups of Negroes and whites. The studies to be reviewed in this section cover a wide range of topics and in some instances are related only by the above broad heading.

Value Systems

The Allport-Vernon Study of Values has been the most frequently employed single measuring instrument (Eagleson & Bell, 1945; Gray, 1947; Milam & Sumner, 1954; Pugh, 1951). Eagleson and Bell (1945) administered this scale to 164 Southern Negroes and compared the results with the original data presented by Allport and Vernon. Subjects in the latter study were primarily from the Northeast. The authors report that the means were not greatly different, but the data were not sufficient to check the significance. Negroes and white females were similar in giving Religious values a top ranking, but differed in that the Negroes scored

lowest on Aesthetic values, while the white females gave this a second place rating. The authors felt that these findings could be explained on the basis of a cultural interpretation relating to the "suffering" of Negroes and the white female's role in society.

Gray (1947) compared these findings with the performance of whites from Peabody College and Florida State University. These subjects also gave Religious values the highest rating with the Social scale being placed second. Both of these scales were significantly higher than all of the others. As compared with the normative group these Southern females also placed a low emphasis upon Aesthetic values. Thus Gray reinforces the conclusion of Eagleson and Bell that low Aesthetic and high Religious scores are a function of culture; this factor seems to apply also to Southern white women. Pugh (1951) administered the Allport-Vernon to several groups (ministers, laymen, nonchurch members) in south Georgia and reported more similarities than differences. Religious values again received top rating by all groups except white men who gave it a rank of five. The ratings of a group of male Negro medical students were similar to those reported above, except that they placed a higher emphasis on the Theoretical values (Milam & Sumner, 1954). These differences seem to be fairly consistent, but once again one is struck by the overall similarities.

Southern Negro and Northern white youth value systems have been analyzed by studying classroom themes (Hughes & Thompson, 1954). For the great majority of values studied a similar amount of interest was reflected in the two groups. Negroes gave a greater emphasis to justice and had a weaker identification with family.

Somner and Killian (1954) investigated value differences by means of an attitude scale wherein the subjects rated the behavior of "a Negro." Negroes want the Negro to be *relatively* forward, passionate, elegant, aggressive, persistent; whereas whites want the Negro to be relatively more witty, jovial, practical, quiet, patient. The emphasis should be on the word *relatively* as the range was restricted by the nature of the scale.

Self-Attitudes

Within recent years there have been a number of studies investigating the young Negro's attitude towards himself. Available evidence indicates that the white or Negro child learns very early about racial differences and that a light skin is to be preferred. Marks (1942) had Negro college students rate each other for attractiveness and found that within their own race there is a preference for light coloring, although extremely light skin was not preferred. A variety of techniques have been employed in evaluating this factor, including the use of white and colored dolls, special puzzles, picture tests, choice of playmates, etc. (Goodman, 1952; Helgerson, 1943; Horowitz, 1947; Koch, 1944; Landreth & Johnson, 1953; Radke & Trager, 1950; Radke, Trager, & Davis, 1949). Goodman (1952) reported finding a marked racial awareness in a group of four-year-old Negroes and whites. Consistent patterns of response to different skin colors has been observed in children of three years of age (Landreth & Johnson, 1953). This study, as well as one by Radke, Sutherland, and Rosenberg (1950), indicates that the development of a negative self-concept has its beginning in the early childhood of the Negro. When compared with whites the Negro is much less positive towards

his own race. Whites tend to express a strong preference for their own race throughout a wide age range, and when a variety of techniques have been employed. In at least one study the proportion of Negroes expressing a preference for Negroes increased with grade level in school (Koch, 1946). Negroes have tended to prefer light skin colors, but a study among Fisk University students reveals that for at least that group the reference scale used in rating skin color is in part a function of the relative colors of the judge and the subject rated (Marks, 1942).

In clinical interviews and autobiographies of "some eighty Negro youths" ranging principally from 17 to 25 years of age (plus some juvenile delinquents and adult patients), Dai (1953) endeavored to determine those problems which are shared with white youth and those peculiar to Negroes. Clinically-oriented, Dai's approach was in terms of role-self-concepts; his interest lay in part in tracing how self-attitudes developed in childhood relate to adult problems.

Dai cites illustrative case material elicited from Negro youth, which could be found in counterpart in most texts on adolescent psychology, in relation to reaction patterns within the primary group, the family: antagonism and open revolt, passive-aggressive reactions, extreme conformity, and so forth. Problems peculiar to Negro children, Dai concludes, are basically human problems, but colored by the fact of being Negro. Caste restrictions bring the following influences to bear on personality development: a preponderance of lower class families with their special (and to the middle class individual, immoral) codes of conduct, broken homes, (consequent) maternal dominance, preoccupation with skin color and other physical features, and an extraordinary stress on social status. The factor most different from being a white youth, i.e., being a youth with dark skin, yields differential reactions in a sense of unworthiness, i.e., absorbing the white's evaluation of dark skin, in developing a sense of indifference—"So what?"—or in identifying with whites to the point of becoming white in judgments of blacks.

Controlled investigations of some of Dai's conclusions are necessary, for his study suffers from the general defects of a posteriori reasoning of many clinical studies and from lack of the precision which even good clinical studies can have. His conclusions do, however, coincide with those reached in the American Council on Education summary of studies (Sutherland, 1942) on "What does it mean for developing personality to be born a Negro?"

Social Perceptions and Attitudes

One study based upon choice of playmate in preschool children gives the racial factor a secondary weighting (Helgerson, 1943). The sex factor was considered the most important single determinant. It may be significant that this particular study was conducted in Minneapolis. In contrast to the Koch study mentioned above, older children chose colored playmates less frequently in both Negro and white groups.

Relatively detailed analysis of children's social perceptions and attitudes may be found in Radke and Trager (1950), Trager and Yarrow (1952), Radke, Trager, and Davis (1949), Bird, Monachesi, and Burdick (1952), and Horowitz (1947). For the most part the results are consistent with the points mentioned above. However, this entire series of studies suffers from many of the methodological errors referred to elsewhere in this paper. In many in-

stances conclusions are based on what appear to be nonsignificant group differences; the race of the examiner is controlled infrequently; social class differences are seldom considered.

Concern with patterns of preference and attitudes toward other racial groups has also been reflected in a series of studies with adults. Studies prior to the period covered by this review suggested that patterns of preference of Negro and white college students were similar. Prothro and Jensen (1952) administered the Grice-Remmers Generalized Attitude Scale to Southern white and Negro college students with results indicating that the preferences of the two groups are not similar. Among other findings the authors report that the attitudes of the Negro towards the white were no more favorable than the reverse. Attitudes of the Negro towards Jews were generally favorable as were those of the whites towards Jews. Gray and Thompson (1953) report different findings on this latter point. On a modified Bogardus Social Distance Scale the Negro subjects voted all groups except their own lower than did the white subjects. The scale was administered to college and high school students as well as randomly selected adults, with white groups being more liberal in all instances.

Interest in the problems of segregation has resulted in several papers comparing the attitudes of Negroes and whites on specific aspects of this issue. Unfortunately, the subjects of these studies have been drawn almost exclusively from high school and college populations. In 1940 Boynton and Mayo (1942) administered a questionnaire to high school students in order to measure attitudinal responses towards the Negro and Negro-white relations. Differences between races were most pronounced with regard to social relations. The authors felt that the differences were becoming greater, primarily because of a shift on the part of the Negro. The questionnaire was administered again in the same schools in 1948 (Mayo & Kinzer, 1950), with the finding of a more favorable attitude towards the Negro by both races. However, the two groups were further from agreement on issues involving interracial relations, and the shift towards a more positive attitude towards the Negro came primarily on items not implying social proximity.

Holmes (1943), working with groups from four colleges reported the Negro more liberal in racial attitudes than the white and the not surprising finding that Negro students in the South were more conservative than those in the North, with white students in the South vastly more conservative than any of the other groups tested. Greenberg, Chase, and Cannon (1957) administered the California F Scale and an integration attitude scale to west Texas high school students. In this instance, the Negro was more authoritarian than the white, but the authoritarian attitudes of either race were not indicative of negative attitudes toward integration.

Scholer (1943) had eighth grade students in Louisiana indicate their degree of approval of racial segregation in hypothetical case situations. Total scores failed to reflect a significant difference between Negroes and whites, although the authors report significant differences in 7 of the 11 situations presented. But the authors conclude that race similarities were more frequent than differences and there were no points of complete contradiction between the two groups. Preferences for segregation

exceeded those for nonsegregation among both whites and Negroes, but to a greater degree among whites.

In a doctoral study Banks (1950) used a slightly different approach and had Ohio Negroes rank 96 different situations as to the degree of resentment experienced. The results were compared with Myrdal's (1944) rank order of discriminations: (1) economic, (2) legal, (3) political, (4) access to public services, facilities and funds, (5) courtesies, (6) sex retions and intermarriage. Myrdal feels that these are reversed for whites. However, the subjects in this study gave the same rankings except that economic discrimination dropped to third place. Thus, it seems that for at least the population employed in this study white and Negro concerns are very similar. In a study of the influence of discrimination on minority group members in New York, a larger percentage of both Negroes and Jews reported experiencing discrimination more frequently than did Catholics or Protestants (Saenger & Gordon, 1950). When compared with the other ethnic groups studied the Negroes particularly felt discriminated against in terms of job opportunities.

One final study dealing directly with attitudes towards segregation and schools should be mentioned. Turman and Holtzman (1955) surveyed a group of teachers in Texas on this question and found that only 4% of the whites and 1% of the Negroes held out consistently for segregation at all levels of education. Forty-four percent of the whites and 57% of the Negroes expressed complete approval of mixed classes in public schools.

Attitudes toward other specific situations have been evaluated in several instances. Clarke and Campbell (1955) had junior high school students estimate their Negro classmates' performance on objective tests. Negroes were significantly more accurate in estimating their fellow Negroes' scores. The estimates of the whites were below those of the Negroes. The authors interpreted the data as reflecting a white stereotype of low Negro ability. One manner in which this stereotype gets reinforced may be shown in an analysis of magazine pictures (Shuey, King, & Griffith, 1953). Pictures of Negroes in leading popular magazines were judged on the basis of the socioeconomic level thought to be reflected. Negroes were portrayed in a less favorable light than in reality and actually appeared in only ½ of 1% of the total pictures.

In Flint, Michigan, students were asked to predict teachers' attitudes toward Negro children (Amos, 1952). In this instance the whites were more accurate than the Negroes who showed more prejudice, stronger feelings of rejection, and more consciousness of race. However, once again when social class was considered there was some evidence that this factor was more important than race in determining pupils' attitudes towards teachers. In a relatively unsophisticated study by Lewis and Biber (1951) Negro children when given an opportunity expressed preference for white teachers; but those who had had a Negro teacher inclined towards choosing a Negro teacher.

Attitudes towards the Negro tend to operate in a number of experimental situations and certainly have influenced the results obtained in the studies reported above. In addition to studies involving the use of psychological tests, the color of the investigator has been shown to be a relevant variable in such diverse

things as work with GSR (Rankin & Campbell, 1955) and identification tests with young children (Trent, 1954).

Overview of Values and Attitudes

Insofar as generalizations can be made from sheer weight, the evidence points to similarities in the value systems of whites and Negroes. Differences in self-concepts are marked, however, in that being a white person in a white society appears to mean little in respect to the development of self-concepts, whereas being a Negro in a white society seems to be one of the most important factors in such development.

The quality of the studies reviewed in this section has varied considerably. Values and attitudes are difficult to quantify, subject to regional variation and other factors which make generalization difficult. One obvious step in clarifying group differences requires the study of populations other than those drawn from schools and colleges.

Vocational Interests

This is another area in which limited research has been reported. A PhD thesis by Hartshorn[4] purportedly discovered interest scores on the Strong Vocational Interest Blank to be different for lawyers, physicians, and life insurance agents of the two groups, in that white men of the standardization group have higher mean interest and occupational level scores, and also more masculine interests, than do their Negro counterparts, and fewer specific likes than do Negroes. Strong (1952) who reviews Hartshorn's work at some length, disputes the latter's conclusions in part, stating the two groups compared are not equivalent (though the major reason Strong offers is not sufficient reason for his contention). Supporting Strong to a slight degree in his denial of the contention that interests of Negro and white professional persons really differ is his research on medical school seniors of both races (Strong, 1955); the differences in occupational and specialization scales are in most cases relatively small. Yet out of 14 occupational scales one is significantly different and six are very significantly different for whites and Negroes, while three out of five specialization scales differ at least at the 5% level of confidence; Strong's concluding remark in his analysis of Hartshorn's data could conceivably do without the qualifying conditional clauses: "If Negroes are really different from whites in their interests, which we are not ready to accept as a fact, then their interests must be measured from a Negro, not a white, point of reference" (Strong, 1955).

Milam and Sumner (1954) have provided further information on the spread and intensity of vocational interests of first year Negro medical students. They report that low vocational interest intensity in Negroes has been a consistent finding with the Strong Vocational Interest Blank and the controversy turns only on why it is lower. After comparing performance on the Strong with academic grades, the authors conclude that there is some evidence to the effect that high scholastic ability is correlated with more intense physician interest as well as more intense and/or diversified nonphysician interests.

Gray (1944a) inquired directly of some 800 Negro children (first to sixth grades) as to what their vocational preferences were. Results were

[4] Known to the writers only through Strong's review.

compared with the responses of white children obtained six years earlier by Boynton. The Negro and white females voiced similar preferences, but the Negro male was more interested in professional occupations than the white male. The median occupational level chosen by the Negro on a five-point scale was one point higher than the white child's. Data were not presented to the extent that statistical significance could be evaluated, but it appears that the Negro children were certainly less realistic. Several studies contrasting Negro and white occupational patterns clearly point to the fact that jobs of higher quality generally are reserved for whites (Keenan & Kerr, 1952; Mundy, 1949; Turner, 1954).

Social Structure

Leadership

Research on comparative qualities of leadership has been extremely limited and, for the most part, has been concerned with the evaluation of miscellaneous factors. Dexter and Stein (1955) administered various temperament tests to leaders and nonleaders in a white and Negro college, with the finding that differences within each group were greater than differences between groups. Some indirect information of a comparative nature is available from studies of *Who's Who in Colored America* (Monahan & Monahan, 1956; Valien & Horton, 1954). Relatively speaking it seems that the Negro female leader receives more recognition than the white female and achieves distinction at a younger age than does the white person of either sex.

Family Organization

In this section various recent miscellaneous studies relating to Negro and white family practices will be reviewed. For comprehensive reviews the reader is referred to the general references listed in the beginning of this article.

The dynamics of the Negro family have been discussed in some detail (Davis & Havighurst, 1946; Frazier, 1939, 1957; Myrdal, 1944). For the most part, the emphasis has been placed upon family disorganization. Statistics have been quoted from the 1950 census to point out the extent to which disrupting socioeconomic and cultural factors are operating (Ginzberg, 1956). At that time one-third of the Negro women who had been married were divorced or separated from their husbands as compared with only one-fifth of the white women. Thirty-five percent of the Negro mothers under 45 were employed versus 19% of the white women in this same age group. In a theoretical discussion of the problems and needs of Negro youth, Frazier (1950) states that the male parent is absent in about 20% of the Negro homes. These statistics would appear to be relatively significant and confirm the existence of greater family disorganization among Negroes, as in general is the case with lower socioeconomic groups (Hollingshead, 1953). If there is any validity to currently held theories of personality development, we would expect some distinctive group personality differences as a reflection of the above facts.

Several studies on child-rearing practices among Negroes and whites have been referred to under the section on intelligence. Davis and Havighurst (1946), after studying groups of mothers from lower and middle classes in Chicago, concluded that essentially the same types of differences prevailed between middle and lower class whites as between middle and lower class Negroes. The

major exception was that Negro mothers were described as more permissive than whites in feeding and weaning, but more rigorous in toilet training. Aside from this the authors conclude that the most striking thing about the study was that the Negro and white middle and lower classes were so much alike.

There have been several comparative investigations of fertility rates. After studying the 1940 census figures, Lee and Lee (1952) state that the pattern of Negro fertility is remarkably similar to that of whites. Within both races fertility declines as socioeconomic level goes up. As with whites, Negro fertility is lower outside of the South. The authors conclude that the patterns of the Negro most closely approach those of the white in those areas where he shares most freely in the general culture. The rate is approximately the same in urban areas; any differences that do appear occur in rural sections. Valien and Vaughn (1951) and Tietze and Lewit (1953), investigated birth control practices in two Southern communities and came up with expected findings: favorable attitudes were correlated with urban birth, education, working mothers, etc.

One investigation of factors entering into mate selection was reported during this period (Sussman & Yeager, 1950). By means of a questionnaire Negro and white college students were rated on 19 different factors. The ratings for the two groups were similar on 14 of these factors, indicating considerable agreement in the qualities thought desirable in a prospective mate. The amount of agreement was surprising since the Negro and white groups were drawn from different sections of the country with differing religious backgrounds. Differences which did appear to be significant are the whites' high rating of "insight and understanding," and the low rating of "good health." For the Negro group these two factors had standings opposite to those of the whites. It is not too remarkable that the health factor is of concern to Southern Negroes as their general level of health is considerably below that of the whites.

Of incidental interest under this section are several studies of Negro-white marriages (Cash, 1956; Golden, 1953, 1954). The subjects of all three of these investigations lived in Philadelphia, and thus generalizations are limited; but the following findings were reported. In most such marriages the male is the Negro; more than 50% of both parties had been previously married; they have few children and tend to marry late; most of the Negroes were Negroid in appearance.

Summarizing the literature on the Negro family dealing with descriptive, theoretical, and quantifiable comparisons with white families, we are led to believe that many of the differences that have been reported can be accounted for in terms of socioeconomic class differences.

Emotional Disturbances and Mental Illness

Attempts to assess the relative emotional stability of Negroes and whites have ranged from the use of paper and pencil tests of neuroticism to comparisons of first admissions to state hospitals. The test approach has not made a significant contribution to a clearer understanding of differences between the two groups, primarily because there are few tests available which have been standardized on Negro populations. In an ex-

perimental study Roberts (1944) has shown that measures of attitudes, adjustment, and personality yield results dependent upon the cultural and racial group in which the measure is made; thus extreme caution is necessary when applying popular American tests to Negroes. It is also recognized that the use of hospital statistics is open to a number of errors.

Attempts to measure Negro-white differences in neuroticism have been limited primarily to college populations. The studies reported have been contradictory. Heyman (1945) has observed that there is a tendency to ignore psychoneurotic behavior in the Negro, even though the observed symptoms are very similar to those in whites. Boykin (1957) administered the Bell Adjustment Inventory to college freshmen over a four-year period and found 25% "poorly adjusted." His findings were compared with the original standardization groups; of his two "poorly adjusted" groups, the Negro group was more maladjusted. It was not possible to tell from the statistics presented if this difference was statistically significant. The Bernreuter Personality Inventory has been administered to Negro groups on at least two occasions. Wheatley and Sumner (1946) found no differences between college student performance and the original norms. They concluded that the most neurotic scores were obtained from the lowest socioeconomic classes. Sumner (1948) on the other hand found no relationship between Bernreuter scores and the socioeconomic status of Negro college women. The California Test of Personality has been administered to second and third grade Negro and white groups without obtaining striking differences, although the author concluded that the minority group tends to feel persecuted (Engle, 1945).

Rowntree (1943) studied discharges from the Service for psychiatric reasons in the period 1941–43 and concluded that a diagnosis of psychoneurosis occurs four times as frequently among Negroes as among whites. The sampling in this study leaves something to be desired; and such factors as geographical representation among the two groups were not controlled. In a study of 105 consecutive discharges from the Navy a diagnosis of psychoneurosis occurred approximately three times as frequently among Negroes as among whites (Hunt, 1947). The author, however, checked these results with four other stations and concluded that the prevalence of psychoneurosis among Negroes is lower than among whites. Gardner and Aaron (1946) reviewed consecutive admissions to the psychiatric ward of a Naval hospital and concluded that whites were more prone to a psychoneurotic reaction than Negroes. In seeming contradiction to these conclusions Ripley and Wolf (1947) reported that minor psychiatric illness occurred two and one-half times more frequently among Negro troops overseas than among comparable white troops. These results were felt to be due in part to lower standards of acceptance for service among Negroes. From these studies we are left wondering whether psychoneurosis actually is more extensive among Negroes or if standards of judging psychoneurosis differ for whites and Negroes.

The picture is clearer when serious mental illness is considered. If first admissions to state hospitals are taken as a measure of relative incidence of psychoses, the occurrence is approximately twice as great among

Negroes (Frumkin, 1954; Ivins, 1950; Malzberg, 1940; Wilson & Lantz, 1957). This ratio seems to hold over a number of years and in widely separated states, with the reported range being from one-and-one-half to four-and-one-half times as great among Negroes. The only contradiction found was in a study by McLean (1949) who reported that first admissions to Illinois State Hospitals were no greater in the Negro than in the white. In explaining this finding he made note of the fact that Illinois is well established as an integrated state.

Even when allowances are made for the fact that whites are more likely to be able to afford private hospitalization, it seems apparent that the relative incidence of psychoses among Negroes is significantly higher. There are also some consistent differences in terms of psychiatric diagnosis and symptoms. Negroes have a higher relative incidence of schizophrenia, paresis, and alcoholic psychosis (Frumkin, 1954; Ivins, 1950; Malzberg, 1940, 1953; Ripley & Wolf, 1947). The only report contradicting this general conclusion, at least in part, is that of Wilson and Lantz (1957), who found that the higher rate of first admissions among Negroes in Virginia was not due to alcoholism, epilepsy, or mental deficiency. They concluded that the difference was due to senile psychoses (although there was a greater number of whites over 65), arteriosclerotic dementia, and schizophrenia.

Other differences in symptomatology have been noted. Psychosomatic disorders are found more frequently among whites, and in some instances —e.g., peptic ulcers—the rate is as much as ten times greater (Rowntree, 1943). The suicide rate seems to differ according to geographical location. McLean (1949) reports that Negroes in the South have a rate only one-fourth that of the whites, whereas the ratio is almost equal in the North. She advanced the explanation that the higher rate among Northern Negroes was due to "ambition." There are also differences in "acting out" behavior which will be discussed in more detail below. In his studies of New York State Hospital, Malzberg (1956a) found that the admission rate for Negroes to hospitals for the criminally insane was four-and-one-half times that of whites.

A number of hypotheses have been advanced to explain the higher incidence of emotional difficulties among Negroes. Various factors associated with socioeconomic status have been of major interest in this regard. Pasamanick, Knoblock, and Lilienfeld (1956) conclude that there is a

> positive and probably etiologic relationship between socioeconomic status and prenatal and paranatal abnormalities which in turn are related to retarded behavioral development and certain NP disorders such as cerebral palsy, mental deficiency, and behavior disorders.

Negroes were reported as having a much greater number of prematurities and complications than white controls with the overall incidence of abnormal conditions being almost twice as great in the Negro. This ratio held up when socioeconomic status was taken into consideration, but since many Negroes were below the lowest white, socioeconomic status was assigned greater weight than any "racial" factors. Hollingshead and Redlich (1953) found in their New Haven study that class is directly related to mental disorder and that schizophrenia occurs most frequently in lower classes. This is one of the psy-

chiatric categories in which the Negro has consistently been reported as having a high incidence.

Other factors felt to contribute to emotional illness in the Negro have centered in family structure, economic problems, and prejudice. Gardner and Aaron (1946) studied childhood and adolescent adjustment and found many similarities between white and Negro psychiatric casualties in a Naval hospital. However, the Negroes were more likely to have an unstable or "broken" home background, and more likely to have been eneuretic in childhood. Ellis and Beechley (1950) compared Negro and white children seen at a child guidance center on a number of variables and concluded that the Negro children were more disturbed, of lower socioeconomic status, came more often from broken homes, and were less responsive to treatment. Racial conflict was found in 32 of 45 consecutive Negro clients of a child guidance clinic and was felt by the investigator to be a contributing factor in the emotional problems of the children (Verin, 1944). An attempt to evaluate cultural factors in mental illness is reported by Cervantes (1954) who studied 30 Northern born Negro patients and 30 "average" whites. Socioeconomic differences between the groups were felt to be important, but the design of the study was such that the differences were not clearly demonstrated. In studying first admissions in state hospitals Malzberg (1956b) reported a higher rate among New York Negro males than among foreign born Negro males. Prejudice was suggested as an explanation for this difference; but the same ratio did not hold for females, a fact which weakens his argument. In another study of first admissions to state hospitals Malzberg (1956a) reports that for the most part Negroes tend to follow the overall pattern in that single, divorced, or widowed individuals have higher rates than married.

Several papers have appeared describing the difficulty in treating psychiatric problems in the Negro as opposed to whites (Adams, 1950; Harms, Kobler, & Sweeney, 1945; Heine, 1950; Kennedy, 1952; St. Clair, 1951). In describing the Negro the following features are frequently mentioned: concern with race consciousness, tendency to act out, hostility as a dominant problem, distrustful, self-hating, strong prestige needs, difficulty in establishing rapport. The referenecs cited are not experimental studies but clinical impressions that are frequently reported and of unknown validity. But there have been several attempts to quantify Negro-white differences in response to treatment. Most of these studies reflect a greater likelihood of improvement in the white. Blassingille (1955) compared the rehabilitation of 70 Negro leucotomy patients at a VA hospital with that of similar patients reported in the literature and considered the results similar. Cultural, social, and economic factors peculiar to the southern Negro were felt not to be significant in rehabilitation of this type patient. In other studies whites have been found to respond more favorably to treatment of varying kinds.

In a review of 455 cases treated with electroshock therapy, the whites were found to be significantly more responsive when the criteria employed were rate of discharge from the hospital and absence of relapse (Palmer, Sprang, & Hans, 1951). Kahn, Buchmueller, and Gildea (1951) describe a program of group therapy for parents of behavior problem chil-

dren which failed when employed with Negroes where it had been considered successful with white parents. The adjustment of Negro and white schizophrenics was studied by Hewlett (1946) with a conclusion that the whites were making the better adjustment. She concluded that economic, financial, occupational, and familial problems were the factors most often related to the adjustment made and that these factors were most severe among Negroes. Ellis and Beechley (1950) feel that some of the same factors were operating to cause a poorer response to treatment among Negro children seen in a child guidance clinic.

For the most part these attempts to evaluate objectively the response to treatment have resulted in conclusions in agreement with clinical impressions reported in the literature. More definite evidence could not be expected in terms of the absence of well accepted criteria of improvement as well as the many extraneous variables operating simultaneously. In evaluating research pertaining to psychiatric disorders among Negroes, Schermerhorn (1956) comments that investigators have frequently been in error in assuming that the Negro and white populations are alike except for the variable in question and that the experimental meaning of these variables is culturally equivalent in both groups. By and large, also, treatment has been offered by white therapists to Negro patients. What effect this factor has in determining results has not been assessed.

On the basis of the evidence available it does appear that the Negroes more frequently experience psychiatric difficulties, particularly of a severe nature. For the most part the specific forms and course appear to be similar with the differences attributable at least in part to identifiable socioeconomic and cultural factors. Whether genetic endowment accounts for the unknown variance is difficult to tell. However, the trend in recent years to find many factors of previously unquestioned hereditary nature attributable to experience makes it also reasonable to assume environmental causation here. Thus, the facts concerning emotional disturbances are fairly clear, but their interpretation is far from being settled.

Crime and Delinquency

Statistics reporting the incidence of crimes and delinquency have consistently shown a higher prevalence among Negroes. The few studies that have appeared in the literature during this period seem not to have been particularly concerned with the question of incidence per se, but have explored possible causative differences between the two races and differential treatment by police officials and the courts.

On the basis of a questionnaire completed by Philadelphia policemen, Kephart (1954) concluded that both white and Negro patrolmen are more strict with Negro offenders. He also noted that Negro offenders more frequently resisted arrest and therefore reinforced the attitude of the officers. Moses (1947) studied crime rates in Baltimore within four socioeconomically equated areas and found in two of them that the Negro rate was higher. He did not find evidence that the Negro offender is convicted more readily. The pattern of offense was similar except that crimes involving loss of life were concentrated among Negroes. The author had some reservations about the attempt to equate the areas in terms of socio-

economic status and found that as a whole the whites had been settled much longer. A greater percentage owned their homes, etc.

Pollack (1944) reviewed Negro and white admissions to Pennsylvania State prisons in 1941–42, being concerned only with subjects who were aged 50 or over. The Negro was charged more frequently with aggressive assault, criminal homicide, and liquor law violations. Pollack reported that the decrease of incidence of various crimes at age 50 and over was comparable in the two groups. The significance of the difference reported could not be evaluated as the data were not reported in the article.

On the basis of statistics compiled in Philadelphia in 1948, Diggs (1950) found several significant differences in both offenses and disposal of offenders between children of the two races. A smaller number of Negro children were dismissed or discharged and a larger number were institutionalized or referred to a criminal court. The Negro was less likely to be referred to private agencies for treatment as opposed to public agencies. The leading offense for Negro boys was that of taking the property of another as opposed to carelessness or mischief for non-Negro boys. Sexual offenses led for Negro girls, while white girls came to the attention of the court most frequently for running away from home. Diggs reported that only one-fourth of the Negro delinquents had both natural parents in the home, but the comparable figure for whites was not reported.

Axelrad (1952) studied the records of 300 institutionalized delinquents in New York City. He concluded that in comparison with whites, Negro children were committed younger, for less serious offenses, with fewer previous court appearances; they came from more unstable homes and homes with a different kind of pathology than that of white delinquents.

The Negro crime rate seems to be consistently higher when a single overall breakdown is made by race. On the basis of recent comparative studies it is difficult to make any judgment as to the relative weights of heredity and environment in this area. It is the writers' opinion that only environmental factors are responsible for differential criminal or delinquent behavior. But if we were to depend on the studies reported here, we must frankly admit we should have little basis for the opinion.

Summary and Concluding Remarks

In psychophysical and psychomotor functions, differences appear between whites and Negroes which may not be accounted for by differential environmental conditions. However, a tendency is present in the literature to indicate that most differences of this nature may be leveled off when social and economic variables are controlled. Intelligence differences reported in the period covered by this review are in the same direction as those seen previously, although infant and young child comparisons suggest greater similarities between Negroes and whites in the early years. Educational achievements of Negroes relative to whites follow the pattern of intellectual differences.

In temperament ("personality") studies Rorschach, TAT, PAT, and P-F Study differences are found, but once again there is insufficient evidence to determine the relative contributions of genetic constitution and experience. At least in those reac-

tions which indicate responses to a dominant-group culture, experience seems to be the major if not sole determinant. Overall, likenesses in psychodynamics appear more extensive than differences. Some, especially paper-and-pencil, tests imply that Negroes are more neurotic than whites, although these tests may be limited for cross-barrier comparisons.

Religious values are ranked first by Negroes, a condition holding generally for white females also, but not for white males. Self-concepts seem to suffer in the Negro subculture in contrast to those of whites. Social perceptions correspondingly differ from one group to the other, with, however, a number of likenesses which may not fit into the stereotype the social scientist holds, for example, in attitudes toward segregation where some Negroes maintain expressed attitudes very much like those of the white majority, or in rank-order of attitudes toward discrimination. Vocational interests may be somewhat similar though less intense (by Strong's criteria) among Negroes; Negro children may be less realistic.

As might be expected, in the areas of psychological functioning most closely related to the sociological, social class differences show up more clearly as bases of differentiation between the two groups. Leadership, family life, child-rearing practices, fertility, and mate selection all seem to conform to social structure rather than to racial lines per se.

Mental illness is considerably more prevalent among Negroes than whites even though specific symptoms like psychosomatic disorders are more frequent in the white group. For various reasons advanced in the body of this article treatment of psychiatric disturbances is more difficult with Negroes than with whites. Rates of delinquency and crime are reported to be higher for Negroes, especially in regard to violent crimes. Most of the evidence indicates that Negroes are given less adequate treatment at the hands of officers of the law and courts.

It is clear from the foregoing review that (*a*) there are still wide differences between Negro and white in many areas of psychological functioning and (*b*) a number of differences attributed in times past to heredity have been shown to be the result of social class determination. It is not clear whether some differences adumbrated here, specifically in the intelligence and temperament realms, are genetically based or not. We agree with Garrett in his Foreword to Shuey's (1958) book, that there are some wholly well-meaning persons who hold that " . . . racial differences ought not to be found; or if found, should immediately be explained away as being somehow immoral and reprehensible." Nevertheless, we are not satisfied that either those who like Garrett believe that genetic differences exist in psychological functions or those who maintain that no such differences can be found have succeeded in establishing their position. Most students in the period of this review have leaned to environmentalist explanations. But this concurrence of judgment may be the result of a possibly unjustified extrapolation from Point b above. We do not agree with Garrett when he says in the same Foreword: "The honest psychologist, like any true scientist, has no preconceived racial bias." As clinical psychologists we are convinced that scarcely anyone undertakes investigation in this field with-

out preconceived biases. We frankly are environmentalist in our bias; but we also hope that we are "honest psychologists" enough to recognize that many research results can yet be interpreted from an hereditarian viewpoint without doing violence to them.

As a last note, from our survey we have come to the conclusion that research within the United States, to which we have limited this review, must be supplemented by investigations between American and other cultures and within other cultures where caste differences are relatively nonexistent.

BIBLIOGRAPHY

ABEL, THEODORA M. Dominant behavior of institutionalized Negro girls: An experimental study. *Amer. J. ment. Defic.*, 1943, **47**, 429–436.

ABEL, THEODORA M. Responses of Negro and white morons to the Thematic Apperception Test. *Amer. J. ment. Defic.*, 1945, **49**, 463–468.

ABEL, THEODORA M., PIOTROWSKI, Z., & STONE, G. Responses of Negro and white morons to the Rorschach test. *Amer. J. ment Defic.*, 1944, **48**, 253–257.

ADAMS, W. A. The Negro patient in psychiatric treatment. *Amer. J. Orthopsychiat.*, 1950, **20**, 305–310.

ALPER, THELMA G., & BORING, E. G. Intelligence test scores of northern and southern white and Negro recruits in 1918. *J. abnorm soc. Psychol.*, 1944, **39**, 471–474.

ALTUS, W. D. The validity of the Terman vocabulary for Army illiterates. *J. consult. Psychol.*, 1946, **10**, 268–276.

ALTUS, W. D., & BELL, H. M. The validity of a general information test for certain groups of Army illiterates. *J. consult. Psychol.*, 1947, **11**, 120–132.

ALTUS, W. D., & CLARK, J. H. Some sectional differences among Negro and white illiterate soldiers. *J. soc. Psychol.*, 1949, **30**, 97–104.

AMOS, R. T. The accuracy of Negro and white children's predictions of teachers' attitudes toward Negro students. *J. Negro Educ.*, 1952, **21**, 125–135.

ANASTASI, ANNE, & D'ANGELO, RITA Y. A comparison of Negro and white preschool children in language development and Goodenough Draw-a-Man I.Q. *J. genet. Psychol.*, 1952, **81**, 147–165.

ANASTASI, ANNE, & FOLEY, J. P., JR. *Differential psychology, individual and group differences in behavior.* (Rev. ed.) New York: Macmillan, 1949.

ANDERSON, W. E. The personality characteristics of 153 Negro pupils, Dunbar High School, Okmulgee, Oklahoma. *J. Negro Educ.*, 1947, **16**, 44–48.

AULD, F., JR. Influence of social class on personality test responses. *Psychol. Bull.*, 1952, **49**, 318–332.

AXELRAD, S. Negro and white male institutionalized delinquents. *Amer. J. Sociol.*, 1952, **57**, 569–574.

BANKS, W. S., JR. The rank order of sensitivity to discrimination of the Negroes in Columbus, Ohio. *Ohio State U., Abstr. doctoral Dissertations. 1948–49.* 1950, No. 60, 7–13.

BEIER, E. G. The effect of induced anxiety on flexibility of intellectual functioning. *Psychol. Monogr.*, 1951, **65**(9, Whole No. 326).

BENDIX, R., & LIPSET, S. M. *Class, status, and power: A reader in social stratification.* Glencoe, Ill.: Free Press, 1953.

BENEDICT, RUTH, & WELTFISH, G. *The races of mankind.* (Public Affairs Pamphlet No. 85) New York: Publ. Affairs Comm., 1943.

BIRD, C., MONACHESI, E. D., & BURDICK, H. Infiltration and the attitudes of white and Negro parents and children. *J. abnorm. soc. Psychol.*, 1952, **47**, 688–699.

BLASSINGILLE, B. Rehabilitation of Negro post-leucotomy patients. *J. nerv. ment. Dis.*, 1955, **121**, 527–534.

BOYD, G. F. The levels of aspiration of white and Negro children in a non-segregated elementary school. *J. soc. Psychol.*, 1952, **36**, 191–196.

BOYD, W. C. *Genetics and the races of man: An introduction to modern physical anthropology.* Boston: Little, Brown, 1950.

BOYKIN, L. L. The reading performance of some Negro college students. *J. Negro Educ.*, 1955, **24**, 435–441.

BOYKIN, L. L. The adjustment of 2,078 Negro students. *J. Negro Educ.*, 1957, **26**, 75–79.

BOYNTON, P. L., & MAYO, G. D. A comparison of certain attitudinal responses of white and Negro high school students. *J. Negro Educ.*, 1942, **11**, 487–494.

BRADLEY, GLADYCE H. A review of educational problems based on military selection and classification data in World War II. *J. educ. Res.*, 1949, **43**, 161–174.

BROWN, ESTELLE W., LYON, R. A., & ANDERSON, NENA A. Courses of prematurity: Influence of maternal illness on incidence of prematurity; employment of new criterion of prematurity for negro race. *AMA J. Dis. Child.*, 1945, **70**, 314.

BROWN, F. An experimental and critical study of the intelligence of Negro and white kindergarten children. *J. genet. Psychol.*, 1944, **65**, 161–175.

BRUNSCHWIG, LILLY. Some personality adjustments of Negro college students as indicated by tests and ratings. *Psychol. Bull.*, 1942, **39**, 578. (Abstract)

BULLOCK, H. A. A comparison of the academic achievements of white and Negro high school graduates. *J. educ. Res.*, 1950, **44**, 179–192.

CALDWELL, M. B. An analysis of responses of a southern urban Negro population to items on the Wechsler Intelligence Scale for Children. Unpublished doctoral dissertation, Penn. State College, 1954.

CANADY, H. G. The problem of equating the environment of Negro-white groups for intelligence testing in comparative studies. *J. soc. Psychol.*, 1943, **17**, 3–15. (a)

CANADY, H. G. A study of sex differences in intelligence test scores among 1306 Negro college freshmen. *J. Negro Educ.*, 1943, **12**, 167–172. (b)

CANADY, H. G. The psychology of the Negro. In P. L. Harriman (Ed.), *The encyclopedia of psychology*. New York: Philosophical Library, 1946. Pp. 407–416.

CASH, E., JR. A study of Negro-white marriages in the Philadelphia area. *Dissertation Abstr.*, 1956, **16**, 1006–1007.

CATTELL, R. B. A culture-free intelligence test: I. *J. educ. Psychol.*, 1940, **31**, 161–180.

CATTELL, R. B. Classical and standard score IQ standardization of the I.P.A.T. Culture-Free Intelligence Scale 2. *J. consult. Psychol.*, 1951, **15**, 154–159.

CATTELL, R. B., FEINGOLD, S. A., & SARASON, S. B. The Culture-Free Intelligence Test: II. Evaluation of cultural influence on test performance. *J. educ. Psychol.*, 1941. **32**, 81–100.

CERVANTES, ESTHER L. A study of cultural factors affecting mental illness. *Smith Coll. Stud. soc. Work*, 1954, **24**, 62–63.

CHAPMAN, W. P. Measurements of pain sensitivity in normal control subjects and in psychoneurotic patients. *Psychosom. Med.*, 1944, **6**, 252–257.

CLARKE, R. B., & CAMPBELL, D. T. A demonstration of bias in estimates of Negro ability. *J. abnorm. soc. Psychol.*, 1955, **51**, 585–588.

CODWELL, J. E. Motor function and the hybridity of the American Negro. *J. Negro Educ.*, 1949, **18**, 452–464.

COOK, R. A. Identification and ego defensiveness in thematic apperception. *J. proj. Tech.*, 1953, **17**, 312–319.

COPPINGER, N. W., & AMMONS, R. B. The Full-Range Picture Vocabulary Test: VIII. A normative study of Negro children. *J. clin. Psychol.*, 1952, **8**, 136–140.

COVELL, L. L. Presbyopia: Comparative observations of white and Negro populations. *Amer. J. Ophthal.*, 1950, **33**, 1275–1277.

DAI, B. Some problems of personality development among Negro children. In C. Kluckholm, H. A. Murray, & D. M. Schneider (Eds.), *Personality in nature, society, and culture*. (2nd ed.) New York: Knopf, 1953.

D'ANGELO, RITA Y. A comparison of white and Negro pre-school children in Goodenough I.Q. and language development. Unpublished master's thesis, Fordham Univer., 1950.

DAVENPORT, R. K. Implications of military selection and classification in relation to universal military training. *J. Negro Educ.*, 1946, **15**, 585–594.

DAVIS, A. *Social class influence upon learning*. Cambridge, Mass.: Harvard Univer. Press, 1948.

DAVIS, A., & EELLS, K. Manual for the Davis-Eells Test of General Intelligence on Problem Solving Ability. Yonkers, N. Y.: World Book, 1953.

DAVIS, A., & HAVIGHURST, R. J. Social class and color differences in child rearing. *Amer. sociol. Rev.*, 1946, **11**, 698–710.

DAVIS, J. C. The scatter pattern of a southern Negro group on the Wechsler Bellevue Intelligence Scale. *J. clin. Psychol.*, 1957, **13**, 298–300.

DESTEPHENS, W. P. Are criminals morons? *J. soc. Psychol.*, 1953, **38**, 187–189.

DEXTER, EMILY S., & STEIN, BETTY. The measurement of leadership in white and Negro students. *J. abnorm. soc. Psychol.*, 1955, **51**, 219–221.

DIGGS, MARY H. Some problems and needs of Negro children as revealed by comparative delinquency and crime statistics. *J. Negro Educ.*, 1950, **19**, 290–297.

DOLLARD, J. *Caste and class in a southern town*. (2nd ed.) New York: Harper, 1949.

DUNN, L. C., & DOBZHANSKY, T. G. *Heredity, race, and society*. New York: Penguin, 1946.

EAGLESON, O. W., & BELL, ELEANOR S. The values of Negro women college students. *J. soc. Psychol.*, 1945, **22**, 149–154.

EELLS, K., DAVIS, A., HAVIGHURST, R. J.,

HERRICK, V. E., & TYLER, R. *Intelligence and cultural differences.* Chicago: Univer. Chicago Press, 1951.

ELLIS, A., & BEECHLEY, R. Comparisons of Negro and white children seen at a child guidance clinic. *Psychiat. quart. Suppl.*, 1950, **24**, 93–100.

ENGLE, T. L. Personality adjustments of children belonging to two minority groups. *J. educ. Psychol.*, 1945, **36**, 543–560.

ESPENSCHADE, ANNA. A note on the comparative motor ability of Negro and white tenth grade girls. *Child Develpm.*, 1946, **17**, 245–248.

EVERHART, R. W. The growth and development of Negro and white elementary children with articulatory defects. *Dissertation Abstr.*, 1953, **13**, 335–336.

EYSENCK, H. J., GRANGER, G. W., & BRENGELMANN, J. C. *Perceptual processes and mental illness.* New York: Basic Books, 1957.

FELTON, JEAN S. The Cornell index used as an appraisal of personality by an industrial health service: I. The total score. *Industr. Med.*, 1949, **18**, 133–144.

FERRELL, G. V. Comparative study of sex differences in school achievement of white and Negro children. *J. educ. Res.*, 1949, **43**, 116–121.

FOWLER, W. L. A comparative analysis of pupil performance on conventional and culture-controlled mental tests. In, *Fourteenth Yearbook of the National Council on Measurements Used in Education.* Princeton, N. J.: Educational Testing Serv., 1957.

FRAZIER, E. F. *The Negro family in the United States.* Chicago: Univer. Chicago Press, 1939.

FRAZIER, E. F. Problems and needs of Negro children and youth resulting from family disorganization. *J. Negro Educ.*, 1950, **19**, 269–277.

FRAZIER, E. F. *The Negro in the United States.* New York: Macmillan, 1957.

FRUMKIN, R. M. Race and major mental disorders: A research note. *J. Negro Educ.*, 1954, **23**, 97–98.

FULK, B. E., & HARRELL, T. W. Negro-white army test scores and last school grade. *J. appl. Psychol.*, 1952, **36**, 34–35.

GARDNER, G. E., & AARON, SADIE. The childhood and adolescent adjustment of Negro psychiatric casualties. *Amer. J. Orthopsychiat.*, 1946, **16**, 481–495.

GARRETT, H. E. Comparison of Negro and white recruits on the Army tests given in 1917–1918. *Amer. J. Psychol.*, 1945, **58**, 480–495. (a)

GARRETT, H. E. A note on the intelligence scores of Negroes and whites in 1918. *J. abnorm. soc. Psychol.*, 1945, **40**, 344–346. (b)

GARRETT, H. E. Psychological differences as among races. *Science*, 1945, **101**, 16–17. (c)

GARRETT, H. E. Negro-white differences in mental ability in the United States. *Scient. Mon.*, 1947, **65**, 329–333.

GARTH, T. A. A review of racial psychology. *Psychol. Bull.*, 1925, **22**, 343–364.

GARTH, T. A. *Race psychology: A study of racial mental differences.* New York: McGraw-Hill, 1931.

GILLILAND, A. R. Socio-economic status and race as factors in infant intelligence test scores. *Child Develpm.*, 1951, **22**, 271–273.

GINZBERG, E. *The Negro potential.* New York: Columbia Univer. Press, 1956.

GINZBERG, E., & BRAY, D. W. *The uneducated.* New York: Columbia Univer. Press, 1953.

GOLDEN, J. Characteristics of the Negro-white intermarried in Philadelphia. *Amer. sociol. Rev.*, 1953, **18**, 177–183.

GOLDEN, J. Patterns of Negro-white intermarriage. *Amer. sociol. Rev.*, 1954, **19**, 144–147.

GOODENOUGH, FLORENCE L., & HARRIS, D. B. Studies in the psychology of children's drawings: II. 1928–1949. *Psychol. Bull.*, 1950, **47**, 369–433.

GOODMAN, MARY E. *Race awareness in young children.* Cambridge, Mass.: Addison-Wesley, 1952.

GRAHAM, T. F. Doll play fantasies of Negro and white primary school children. *J. clin. Psychol.*, 1955, **11**, 29–33.

GRAY, J. S., & THOMPSON, A. H. The ethnic prejudices of white and Negro college students. *J. abnorm. soc. Psychol.*, 1953, **48**, 311–313.

GRAY, SUSAN W. The vocational preferences of Negro school children. *J. genet. Psychol.*, 1944, **64**, 239–247. (a)

GRAY, SUSAN W. The wishes of Negro school children. *J. genet. Psychol.*, 1944, **64**, 225–237. (b)

GRAY, SUSAN W. A note on the values of southern college women, white and Negro. *J. soc. Psychol.*, 1947, **25**, 239–241.

GREENBERG, H., CHASE, A. L., & CANNON, T. M., JR. Attitudes of white and Negro high school students in a west Texas town toward school integration. *J. appl. Psychol.*, 1957, **41**, 27–31.

GUILFORD, J. P. The structure of intellect. *Psychol. Bull.*, 1956, **53**, 267–293.

GUILFORD, J. P. Three faces of intellect. *Amer. Psychologist*, 1959, **14**, 469–479.

HAMMER, E. F. Frustration-aggression hypothesis extended to socio-racial areas: Com-

parison of Negro and white children's H-T-P's. *Psychiat. Quart.*, 1953, **27**, 597–607. (a)

HAMMER, E. F. Negro and white children's personality adjustment as revealed by a comparison of their drawings (H-T-P). *J. clin. Psychol.*, 1953, **9**, 7–10. (b)

HAMMER, E. F. Comparison of the performance of Negro children and adolescents on two tests of intelligence, one an emergency scale. *J. genet. Psychol.*, 1954, **84**, 85–93.

HARMS, H. E., KOBLER, F. J., & SWEENEY, F. J. Negro Army psychiatric casualties. *War Med.*, 1945, **7**, 309–312.

HARROWER-ERICKSON, MOLLY R., & STEINER, M. E. *Large scale Rorschach techniques.* Springfield, Ill.: Charles C Thomas, 1945.

HEINE, R. W. The Negro patient in psychotherapy. *J. clin. Psychol.*, 1950, **6**, 373–376.

HELGERSON, EVELYN. The relative significance of race, sex, and facial expression in choice of playmate by the preschool child. *J. Negro Educ.*, 1943, **12**, 617–622.

HELSON, H., & GUILFORD, J. P. The relation of visual sensitivity to the amount of retinal pigmentation. *J. gen. Psychol.*, 1933, **9**, 58–76.

HESS, A. D. Controlling culture influence in mental testing: An experimental test. *J. educ. Res.*, 1955, **49**, 53–58.

HEWLETT, JANE D. A comparison of the adjustment of white and Negro schizophrenics. *Smith Coll. Stud. soc. Work*, 1946, **17**, 108–109.

HEYMAN, DOROTHY. Manifestations of psychoneurosis in Negroes. *Ment. Hyg.*, 1945, **29**, 231–235.

HOFFMAN, M. L., & ALBIZU-MIRANDA, C. Middle class bias in personality testing. *J. abnorm. soc. Psychol.*, 1955, **51**, 150–152.

HOLLINGSHEAD, A. B. Class differences in family stability. In R. Bendix & S. M. Lipset (Eds.), *Class, status and power: A reader in social stratification.* Glencoe, Ill.: Free Press, 1953. Pp. 284–292.

HOLLINGSHEAD, A. B., & REDLICH, F. C. Social stratification and psychiatric disorders. *Amer. sociol. Rev.*, 1953, **18** (2), 163–169.

HOLMES, JEAN B. A study of racial attitudes found in four colleges, including a scale to measure these attitudes. *J. Negro Educ.*, 1943, **12**, 69–75.

HOROWITZ, E. L. Development of attitude toward Negroes. In T. M. Newcomb & E. L. Hartley (Eds.), *Readings in social psychology.* New York: Henry Holt, 1947. Pp. 507–517.

HORST, P. Multiple classification by the method of least squares. *J. clin. Psychol.*, 1956, **12**, 3–16.

HUGHES, J. H., & THOMPSON, G. G. A comparison of the value systems of southern Negro and northern white youth. *J. educ. Psychol.*, 1954, **45**, 300–309.

HUNT, W. A. The relative incidence of psychoneurosis among Negroes. *J. consult. Psychol.*, 1947, **11**, 133–136.

IRWIN, O. C. Infant speech. *Scient. Amer.*, 1949, **181**(3), 22–24.

IVINS, S. P. Psychoses in the Negro: A preliminary study. *Del. State med. J.*, 1950, **22**, 212–213.

JENKINS, M. D. Case studies of Negro children of Binet IQ 160 and above. *J. Negro Educ.*, 1943, **12**, 159–166.

JENKINS, M. D. The upper limit of ability among American Negroes. *Scient. Mon.*, 1948, **66**, 399–401.

JENKINS, M. D. Intellectually superior Negro youth: Problems and needs. *J. Negro Educ.*, 1950, **19**, 322–332.

JENKINS, M. D., & RANDALL, CONSTANCE M. Differential characteristics of superior and unselected Negro college students. *J. soc. Psychol.*, 1948, **27**, 187–202.

KAHN, JANE, BUCHMUELLER, A. D., & GILDEA, MARGARET C. L. Group therapy for parents of behavior problem children in public schools: Failure of the method in a Negro school. *Amer. J. Psychiat.*, 1951, **108**, 351–357.

KARDINER, A., & OVESEY, L. *The mark of oppression: A psychosocial study of the American Negro.* New York: Norton, 1951.

KARON, B. P. *The Negro personality: A rigorous investigation of the effects of culture.* New York: Springer, 1958.

KARPINOS, B. D. Visual acuity of selectees and Army inductees. *Hum. Biol.*, 1944, **16**, 1–14.

KEENAN, V., & KERR, W. A. Unfair employment practices as viewed by private employment counselors. *J. appl. Psychol.*, 1952, **36**, 361–364.

KENNEDY, JANET A. Problems posed in the analysis of Negro patients. *Psychiatry*, 1952, **15**, 313–327.

KEPHART, W. M. The Negro offender: An urban research project. *Amer. J. Sociol.*, 1954, **60**, 46–50.

KLINEBERG, O. *Race differences.* New York: Harper, 1935.

KLINEBERG, O. (Ed.), *Characteristics of the American Negro.* New York: Harper, 1944.

KLUGMAN, S. F. The effect of money incentives versus praise upon the reliability and obtained scores of the revised Stanford-Binet test. *J. gen. Psychol.*, 1944, **30**, 255–269.

KNOBLOCK, HILDA, & PASAMANICK, B. Fur-

ther observation on the behavioral development of Negro children. *J. genet. Psychol.*, 1953, **83**, 137–157.

KNOX, E. O. The Negro as a subject of university research in 1944. *J. Negro Educ.*, 1945, **14**, 182–196.

KNOX, E. O. The Negro as a subject of university research in 1948. *J. Negro Educ.*, 1949, **18**, 489–497.

KNOX, E. O. The Negro as a subject of university research in 1951. *J. Negro Educ.*, 1952, **21**, 484–491.

KOCH, HELEN L. A study of some factors conditioning the social distance between the sexes. *J. soc. Psychol.*, 1944, **20**, 79–107.

KOCH, HELEN L. The social distance between certain racial, nationality, and skin pigmentation groups in selected populations of American school children. *J. genet. Psychol.*, 1946, **68**, 63–95.

KORCHIN, S. J., MITCHELL, H. E., & MELTZOFF, J. A critical evaluation of the Thompson Thematic Apperception Test. *J. proj. Tech.*, 1950, **14**, 445–452.

LANDRETH, CATHERINE, & JOHNSON, BARBARA C. Young children's responses to a picture and inset test designed to reveal reactions to persons of different skin color. *Child Develpm.*, 1953, **24**, 63–80.

LEE, E. S. Negro intelligence and selective migration: A Philadelphia test of the Klineberg hypothesis. *Amer. sociol. Rev.*, 1951, **16**, 227–233.

LEE, E. S., & LEE, ANNE S. The differential fertility of the American Negro. *Amer. sociol. Rev.*, 1952, **17**, 437–447.

LEWIS, CLAUDIA, & BIBER, BARBARA. Reactions of Negro children toward Negro and white teachers. *J. exp. Educ.*, 1951, **20**, 97–104.

LIGHT, B. H. A further test of the Thompson TAT rationale. *J. abnorm. soc. Psychol.*, 1955, **51**, 148–150.

LINDZEY, G. (Ed.) *Handbook of social psychology.* Cambridge, Mass.: Addison-Wesley, 1954.

LONG, H. H. The relative learning capacity of Negroes and whites. *J. Negro Educ.*, 1957, **26**, 121–134.

LOVE, M. I., & BEACH, S. Performance of children on the Davis-Eells Games and other measures of ability. *J. consult. Psychol.*, 1957, **21**, 29–32.

MCCARY, J. L. Reactions to frustrations by some cultural and racial groups. *Personality*, 1951, **1**, 84–102.

MCCARY, J. L. Picture Frustration Study normative data for some cultural and racial groups. *J. clin. Psychol.*, 1956, **12**, 194–195.

MCCARY, J. L., & TRACKTIR, J. Relationship between intelligence and frustration-aggression patterns as shown by two racial groups. *J. clin. Psychol.*, 1957, **13**, 202–204.

MCGRAW, MYRTLE B. A comparative study of a group of southern white and Negro infants. *Genet. psychol. Monogr.*, 1931, **10**, 1–105.

MCGURK, F. C. J. Comparative test scores of Negro and white school children in Richmond, Va. *J. educ. Psychol.*, 1943, **34**, 474–484.

MCGURK, F. C. J. *Comparison of the performance of Negro and white high school seniors on cultural and non-cultural test questions.* Washington, D. C.: Catholic Univer. Amer. Press, 1951. (Microcard)

MCGURK, F. C. J. On white and Negro test performance and socioeconomic factors. *J. abnorm. soc. Psychol.*, 1953, **48**, 448–450. (a)

MCGURK, F. C. J. Socio-economic status and culturally-weighted test scores of Negro subjects. *J. appl. Psychol.*, 1953, **37**, 276–277. (b)

MACHOVER, S. *Cultural and racial variations in patterns of intellect.* New York: Teachers Coll., Columbia Univer., 1943.

MCLEAN, HELEN H. The emotional health of Negroes. *J. Negro Educ.*, 1949, **18**, 283–290.

MACPHEE, H. M., WRIGHT, H. F., & CUMMINGS, S. B. The performance of mentally subnormal rural southern Negroes on the verbal scale of the Bellevue Intelligence Examination. *J. soc. Psychol.*, 1947, **25**, 217–229.

MALZBERG, B. *Social and biological aspects of mental disease.* New York: N. Y. State Hospital Press, 1940.

MALZBERG, B. Mental disease among Negroes in New York State, 1939–41. *Ment. Hyg.*, 1953, **37**, 450–476.

MALZBERG, B. Marital status and mental disease among Negroes in New York State. *J. nerv. ment. Dis.*, 1956, **123**, 457–465. (a)

MALZBERG, B. Mental disease among native and foreign-born Negroes in New York State. *J. Negro Educ.*, 1956, **25**, 175–181. (b)

MARKS, E. S. Factors affecting skin color judgments of Negro college students. *Psychol. Bull.*, 1942, **39**, 577. (Abstract)

MARQUART, DOROTHY, & BAILEY, LOIS L. An evaluation of the Culture-Free Test of Intelligence. *J. genet. Psychol.*, 1955, **86**, 353–358.

MAYO, G. D., & KINZER, J. R. A comparison of the "racial" attitudes of white and Negro high school students in 1940 and 1948. *J. Psychol.*, 1950, **29**, 397–405.

MEREDITH, H. V. Physical growth from birth to two years: I. Stature. A review and synthesis of North American research for the period 1850–1941. *U. Ia. Stud. child Welf.*, 1943, **19**, No. 407.

MICHAEL, W. B. An investigation of the contribution of factors to tests and to their predictive value in two Army Air Force pilot populations. *Amer. Psychologist*, 1947, **2**, 417–418. (Abstract)

MICHELSON, N. Studies in the physical development of Negroes: II. Weight. *Amer. J. phys. Anthrop.*, 1943, **1**, 289–300.

MILAM, A. T., & SUMNER, F. C. Spread and intensity of vocational interests and evaluative attitudes in first year Negro medical students. *J. Psychol.*, 1954, **37**, 31–38.

MINER, J. B. *Intelligence in the United States: A survey—with conclusions for manpower utilization in education and employment.* New York: Springer, 1957.

MONAHAN, T. P., & MONAHAN, ELIZABETH H. Some characteristics of American Negro leaders. *Amer. sociol. Rev.*, 1956, **21**, 589–596.

MONTAGU, M. F. A. *Man's most dangerous myth: The fallacy of race.* (3rd ed.) New York: Harper, 1952.

MOORE, J. E. A comparison of Negro and white pre-school children on a vocabulary test and eye-hand coordination test. *Child Develpm.*, 1942, **13**, 247–252.

MOSES, E. R. Differentials in crime rates between Negroes and whites, based on comparisons of four socio-economically equated areas. *Amer. sociol. Rev.*, 1947, **12**, 411–420.

MUNDY, P. The young Negro worker in Washington, D. C. *J. Negro Educ.*, 1949, **18**, 104–113.

MUSSEN, P. H. Differences between the TAT responses of Negro and white boys. *J. consult. Psychol.*, 1953, **17**, 373–376.

MYRDAL, G. *An American dilemma: The Negro problem and modern democracy.* New York: Harper, 1944. 2 vols.

NEWLAND, T. E., & LAWRENCE, W. C. Chicago non-verbal examination results on an east Tennessee Negro population. *J. clin. Psychol.*, 1953, **9**, 44–47.

NORTH, R. D. The intelligence of American Negroes. In M. M. Turmin (Ed.), *Segregation and desegregation: A digest of recent research.* New York: Anti-defamation League of B'nai B'rith, 1957.

PALMER, D. M., SPRANG, H. E., & HANS, C. L. Electroshock therapy in schizophrenia: A statistical survey of 455 cases. *J. nerv. ment. Dis.*, 1951, **114**, 162–171.

PASAMANICK, B. A comparative study of the behavioral development of Negro infants. *J. genet. Psychol.*, 1946, **69**, 3–44.

PASAMANICK, B., & KNOBLOCH, HILDA. Early language behavior in Negro children and the testing of intelligence. *J. abnorm. soc. Psychol.*, 1955, **50**, 401–402.

PASAMANICK, B., KNOBLOCH, HILDA, & LILIENFELD, A. M. Socioeconomic status and some precursors of neuropsychiatric disorder. *Amer. J. Orthopsychiat.*, 1956, **26**, 594–601.

POLLACK, O. A statistical investigation of the criminality of old age. *J. clin. Psychopath.*, 1944, **5**, 745–767.

PORTNOY, B., & STACEY, C. L. A comparative study of Negro and white subnormals on the children's form of the Rosenzweig P-F Test. *Amer. J. ment. Defic.*, 1954, **59**, 272–278.

PROTHRO, E. T., & JENSEN, J. A. Comparison of some ethnic and religious attitudes of Negro and white college students in the deep South. *Soc. Forces*, 1952, **30**, 426–428.

PROTHRO, E. T., & PERRY, H. T. Group differences in performance on the Meier Art Test. *J. appl. Psychol.*, 1950, **34**, 96–97.

PUGH, T. J. A comparative study of the values of a group of ministers and two groups of laymen. *J. soc. Psychol.*, 1951, **33**, 225–235.

RADKE, MARIAN J., SUTHERLAND, JEAN, & ROSENBERG, PEARL. Racial attitudes of children. *Sociometry*, 1950, **13**, 154–171.

RADKE, MARIAN J., & TRAGER, HELEN G. Children's perceptions of the social roles of Negroes and whites. *J. Psychol.*, 1950, **29**, 3–33.

RADKE, MARIAN J., TRAGER, HELEN G., & DAVIS, HADASSAH. Social perceptions and attitudes of children. *Genet. psychol. Monogr.*, 1949, **40**, 327–447.

RAMSEY, G. V. Sexual growth of Negro and white boys. *Hum. Biol.*, 1950, **22**, 146–149.

RANKIN, R. E., & CAMPBELL, D. T. Galvanic skin response to Negro and white experimenters. *J. abnorm. soc. Psychol.*, 1955, **51**, 30–33.

REITZES, D. C. *Negroes and medicine.* Cambridge, Mass.: Harvard Univer. Press, 1958.

RHOADS, TERESA F., RAPOPORT, M., KENNEDY, RUTH, & STOKES, J. Studies on the growth and development of male children receiving evaporated milk: II, Physical growth, dentition, and intelligence of white and Negro children through the first four years as influenced by vitamin supplements. *J. Pediat.*, 1941, **26**, 415–454.

RIESS, B. F., SCHWARTZ, E. K., & COTTINGHAM, ALICE. An experimental critique of assumptions underlying the Negro version

of the TAT. *J. abnorm. soc. Psychol.*, 1950, **45**, 700–709.

RIPLEY, H. S., & WOLF, S. Mental illness among Negro troops overseas. *Amer. J. Psychiat.*, 1947, **103**, 499–512.

ROBERTS, S. O. The measurement of adjustment of Negro college youth: Personality scales for whites versus criteria intrinsic to Negro groups. Unpublished doctor's dissertation, University of Minn., 1944.

ROBERTS, S. O. Socio-economic status and performance of Negro college women, north and south, on the ACE. *Amer. Psychologist*, 1946, **1**, 253. (Abstract)

ROBERTS, S. O. Socio-economic status and performance on the ACE of Negro freshmen college veterans and non-veterans from the north and the south. *Amer. Psychologist*, 1948, **3**, 266. (Abstract)

ROBERTS, S. O. Socio-economic status and performance over a four-year period on the ACE of Negro college women from the north and south. *Amer. Psychologist*, 1950, **5**, 295. (Abstract)

ROBINSON, MARY LOUISE, & MEENES, M. The relationship between test intelligence of third grade Negro children and the occupations of their parents. *J. Negro Educ.*, 1947, **16**, 136–141.

RODGERS, E. G. The relationship of certain measurable factors in the personal and education backgrounds of two groups of Baltimore Negroes, identified as superior and average in intelligence as fourth grade children, to their educational, social, and economic achievement in adulthood. *Dissertation Abstr.*, 1957, **17**, 1709.

ROWNTREE, L. G. The unfit: How to exclude them. *Psychosom. Med.*, 1943, **5**, 324–330.

SAENGER, G., & GORDON, N. S. The influence of discrimination on minority group members in its relation to attempts to combat discrimination. *J. soc. Psychol.*, 1950, **31**, 95–120.

ST. CLAIR, H. R. Psychiatric interview experiences with Negroes. *Amer. J. Psychiat.*, 1951, **108**, 113–119.

SARASON, S. B., & GLADWIN, T. Psychological and cultural problems in mental subnormality. *Genet. psychol. Monogr.*, 1958, **57**, 3–290.

SCARBOROUGH, B. B. Some mental characteristics of southern colored and white venereal disease patients as measured by the Wechsler-Bellevue Test. *J. soc. Psychol.*, 1956, **43**, 313–321.

SCHERMERHORN, R. A. Psychiatric disorders among Negroes: A sociological note. *Amer. J. Psychiat.*, 1956, **112**, 878–882.

SCHOLER, E. A. Attitudes toward racial segregation in Baton Rouge, Lousiana. *J. soc. Psychol.*, 1943, **18**, 33–53.

SCHWARTZ, E. K., RIESS, B. F., & COTTINGHAM, ALICE. Further critical evaluation of the Negro version of the TAT. *J. proj. Tech.*, 1951, **15**, 394–400.

SCOTT, R. B., CARDOZO, W. W., SMITH, A. D., & DELILLY, M. R. Growth and development of Negro infants: III. Growth during the first year of life as observed in private pediatric practice. *J. Pediat.*, 1950, **37**, 885–893.

SCOTT, R. B., FERGUSON, ANGELLA D., JENKINS, M. E., & CUTLER, F. E. Growth and development of Negro infants: V. Neuromuscular patterns of behavior during the first year of life. *Pediatrics*, 1955, **16**, 24–30.

SHUEY, AUDREY M. *The testing of Negro intelligence.* Lynchburg, Va.: Randolph-Macon Women's College, 1958.

SHUEY, AUDREY M., KING, NANCY, & GRIFFITH, BARBARA. Stereotyping of Negroes and whites: An analysis of magazine pictures. *Publ. opin. Quart.*, 1953, **17**, 281–287.

SOMNER, R., & KILLIAN, L. M. Areas of value differences: I. A method of investigation. II. Negro-white relations. *J. soc. Psychol.*, 1954, **39**, 227–235, 237–244.

STAINBROOK, E., & SIEGEL, P. S. A comparative group Rorschach study of southern Negro and white high school and college students. *J. Psychol.*, 1944, **17**, 107–115.

STRONG, E. K., JR. Interests of Negroes and whites. *J. soc. Psychol.*, 1952, **35**, 139–150.

STRONG, E. K., JR. Are medical specialist interest scales applicable to Negroes? *J. appl. Psychol.*, 1955, **39**, 62–64.

SUMNER, F. C. Neurotic tendency and socio-economic status of Negro college women. *J. soc. Psychol.*, 1948, **28**, 291.

SUNNE, D. Comparative study of white and Negro children. *J. appl. Psychol.*, 1917, **1**, 71–84.

SUSSMAN, M. B., & YEAGER, H. C., JR. Mate selection among Negro and white college students. *Sociol. soc. Res.*, 1950, **35**, 46–49.

SUTHERLAND, R. L. *Color, class, and personality.* Washington, D. C.: Amer. Council on Education, 1942.

THEMAN, VIOLA, & WITTY, P. Case studies and genetic records of two gifted Negroes. *J. Psychol.*, 1943, **15**, 165–181.

THOMPSON, C. E. The Thompson modification of the Thematic Apperception Test. *J. proj. Tech.*, 1949, **13**, 469–478.

THOMPSON, D. C. Teachers in Negro colleges: A sociological analysis. *Dissertation Abstr.*, 1956, **16**, 1531–1532.

THURSTONE, L. L. Primary mental abilities. *Psychometr. Monogr.*, 1938, No. 1.

TIETZE, C., & LEWIT, SARAH. Patterns of family limitations in a rural Negro community. *Amer. sociol. Rev.*, 1953, **18**, 563–564.

TOMKINS, S. S., & MINER, J. B. *The Picture Arrangement Test.* New York: Springer, 1957.

TOUCHSTONE, F. V. A comparative study of Negro and white college students' aggressiveness by means of sentence completion. *Dissertation Abstr.*, 1957, **17**, 1588–1589.

TRAGER, HELEN G., & YARROW, MARIAN R. *They learn what they live.* New York: Harper, 1952.

TRENT, R. D. The color of the investigator as a variable in experimental research with Negro subjects. *J. soc. Psychol.*, 1954, **40**, 281–287.

TURMAN, J. A., & HOLTZMAN, W. H. Attitudes of white and Negro teachers toward nonsegregation in the classroom. *J. soc. Psychol.*, 1955, **42**, 61–70.

TURNER, R. H. Occupational patterns of inequality. *Amer. J. Sociol.*, 1954, **59**, 437–447.

TYLER, LEONA E. *The psychology of human differences.* (2nd ed.) New York: Appleton-Century-Crofts, 1956.

VALIEN, P., & HORTON, CARRELL. Some demographic characteristics of outstanding Negro women. *J. Negro Educ.*, 1954, **23**, 406–420.

VALIEN, P., & VAUGHN, RUTH E. Birth control attitudes and practices of Negro mothers. *Sociol. soc. Res.*, 1951, **35**, 415–421.

VERIN, OLGA. Manifestation of racial conflict in Negro clients of a child guidance clinic. *Smith Coll. Stud. soc. Work*, 1944, **15**, 126–127.

WHEATLEY, L. A., & SUMNER, F. C. Measurement of neurotic tendency in Negro students of music. *J. Psychol.*, 1946, **22**, 247–252.

WILLIAMS, JUDITH, & SCOTT, R. B. Growth and development of Negro infants: IV. Motor development and its relationship to child rearing practices in two groups of Negro infants. *Child Develpm.*, 1953, **24**, 103–121.

WILSON, D. C., & LANTZ, EDNA M. The effect of culture change on the Negro race in Virginia, as indicated by a study of state hospital admissions. *Amer. J. Psychiat.*, 1957, **114**, 25–32.

WINSLOW, C. N., & BRAINERD, J. E. A comparison of the reaction of whites and Negroes to frustration as measured by the Rosenzweig Picture-Frustration Test. *Amer. Psychologist*, 1950, **5**, 297. (Abstract)

WITTY, P. A., & JENKINS, M. D. The case of "B," a gifted Negro girl. *J. soc. Psychol.*, 1935, **6**, 117–124.

WITTY, P., & THEMAN, VIOLA. A follow-up study of the educational attainment of gifted Negroes. *J. educ. Psychol.*, 1943, **34**, 35–47.

WOISIKA, P. H. An evaluation of the dark test. *Amer. intern. Med.*, 1944, **21**, 101–118.

WOODS, R. C., & MARTIN, L. R. Testing in musical education. *Educ. psychol. Measmt.*, 1943, **3**, 29–42.

WOODS, W. A., BOGER, J., & HOLMAN, G. An investigation of revised beta scores (Lindner Gurvitz Scale) among Negro adolescents. *Va. J. Sci.*, 1954, **4**, 321.

WOODS, W. A., & TOAL, R. Subtest disparity of Negro and white groups matched for IQs on the Revised Beta Test. *J. consult. Psychol.*, 1957, **21**, 136–138.

WOODWORTH, R. S. Comparative psychology of races. *Psychol. Bull.*, 1916, **13**, 388–397.

YOUNG, FLORENE M., & BRIGHT, H. A. Results of testing 81 Negro rural juveniles with the Wechsler Intelligence Scale for Children. *J. soc. Psychol.*, 1954, **39**, 219–226.

YULE, G. U., & KENDALL, M. G. *An introduction to the theory of statistics.* (13th ed.) London: Griffin, 1949.

(Early Publication received January 4, 1960)

Psychological Bulletin
1960, Vol. 57, No. 5, 403-415

EMPIRICAL FINDINGS AND THEORETICAL PROBLEMS IN THE USE OF ANXIETY SCALES[1]

IRWIN G. SARASON
University of Washington

In terms of productivity during the past decade, few areas of study in psychology have matched the output of research on scales of anxiety. While the inundation of papers on anxiety has impressed some workers and troubled others, it behooves us to inquire into (*a*) the stimulus value of anxiety scales for psychologists, (*b*) the contribution of research on anxiety to the body of psychological knowledge, and (*c*) the problems for future study raised by this research. It is the purpose of this paper to attempt such an evaluation with emphasis on the relationship of anxiety to stress, learning, intelligence, physiological responses, other personality characteristics, and test taking attitudes. The purport of the paper is not to present a general review of all studies on anxiety in these areas but rather to attempt to abstract from a large literature major trends which seem of present or potential significance.

An attempt such as the present one seems particularly appropriate in view of several recent evaluations of research involving anxiety scales which noted the unreplicability and inconsistencies of certain reported findings in this field (Bendig & Vaughan, 1957; Blake & Mouton, 1959; Farber & Spence, 1956; Jensen, 1958; McClelland, 1958). Frustrating as this state of affairs may be, the present writer will attempt to show that unreplicability is not necessarily attributable to unreliability in the anxiety measuring instruments, but rather, to several "traditional" variables such as characteristics of *Ss* and *Es*, and population and instructional variables which confound with anxiety measures.

[1] The preparation of this paper was facilitated by a grant (M-2397) from the National Institute of Mental Health.

The Stimulus Value for the Psychologist of Anxiety Scales

In view of the centrality of the concept of anxiety in personality theory, it is somewhat surprising that attempts to measure the concept objectively have developed only in recent years. Also, psychologists concerned with personality functioning might well be surprised at the context in which the first widely used anxiety scale was developed. A group of experimental psychologists interested in problems of learning was responsible for the development of Taylor's Manifest Anxiety Scale (MAS) (Farber, 1955; Taylor, 1951, 1953, 1956). The main interest of these researchers in the MAS was in the measurement of Hull's *D* in human *Ss* who were being studied in learning situations.

Whereas the work stemming from the Iowa laboratory was concerned with the relationship of MAS to *D*, other researchers have inquired into the relationship between anxiety measures and a host of varied behaviors and situations (Eichhorn & Tracktir, 1955; Eriksen & Wechsler, 1955; Fiedler, Dodge, Jones, & Hutchins, 1958; Janis, 1955; M. T. Mednick, 1957; Rosenbaum, 1956; Siegal, 1954; Taft, 1957; Westrope, 1953; Wolf,

1955). Motivated by the need for measures of personality relevant to such variables as intellectual performance, reaction to stress, and ability to learn, psychologists seized upon the objective, easily administered MAS. In view of the absence of measures of individual differences in anxiety, the motivation underlying the swift adoptions of the MAS seems clear. However, the criticism of Jenkins and Lykken (1957) that in some research projects involving the MAS the rationale for its use has been lacking seems to be a just one.

The availability of the MAS served to stimulate its use by researchers with varied interests, and it has also encouraged other investigators to construct other measures of anxiety better fitted to their specific needs (Bendig, 1956; Dixon, deMonchaux, & Sandler, 1957; Lykken, 1957; Mandler & Sarason, 1952; Sarason, 1958b; Welsh, 1952, 1956). As a result, measures for specific anxieties such as test anxiety, social anxiety, and anxiety in children are now readily available. There is reason to believe that the various measures of anxiety in current use are not all measuring the same thing (Feldman & Siegel, 1958; Goodstein, 1954; Gordon & Sarason, 1955; Jackson & Bloomberg, 1958; Lauterbach, 1958; Sarason, 1959a; Sinick, 1956; Windle, 1955; Zimet & Brackbill, 1956). An important current problem is the clarification of the similarities and differences among existing anxiety indices. In this connection, Jessor and Hammond (1957), working within the framework of Cronbach and Meehl's (1955) concept of construct validity, have provided some very useful suggestions. It is to be hoped that the days of naive acceptance of the face validity of anxiety scales may be numbered and that more concern will be given to the theoretical bases underlying the use of particular measures of anxiety.

Relationship of Anxiety to Behavior

As has already been indicated, existing studies of anxiety literally defy summary as a unit. However, it is possible to discern trends and problems in certain areas where anxiety scales have been employed, and it is with these that this paper will be concerned.

Anxiety and Stress

Many investigators have studied the reactions of *Ss* differing in scores on anxiety scales to situations posing personal threat or stress for *Ss*. Typically, the stress has been created by means of verbal instructions, e.g., informing *S* he is about to take an intelligence test. Most investigators have assumed that high anxious *Ss* would be more sensitive to implied personal threat than would low anxious *Ss*.

Although some investigators (Cox & Sarason, 1954; Farber & Spence, 1956; Gynther, 1957; Taylor, 1958) have presented evidence not consistent with this assumption, the bulk of the available findings suggest that high anxious *Ss* are affected more detrimentally by motivating conditions or failure reports than are *Ss* lower in the anxiety score distribution (Davidson, Andrews, & Ross, 1956; Gordon & Berlyne, 1954; Korchin & Levine, 1957; Lucas, 1952; Mandler & Sarason, 1952; Nicholson, 1958; Sarason, 1956a, 1957a, 1957b, 1959b, 1959c; Sarason, Mandler, & Craighill, 1952; Sarason & Palola, 1960; Truax & Martin, 1957; Westrope, 1953). Illustrative of this type of study is that of Davidson, Andrews, and Ross (1956) in which three variables were studied:

(*a*) MAS scores, (*b*) reports to *Ss* of levels of failure, and (*c*) speed of presentation of task stimuli. Significant interactions were obtained among all of the variables, and the authors concluded that high anxious *Ss* are more sensitive to experimental stress than are low anxious *Ss*.

In this connection it is interesting to note that high anxious *Ss* have been found to be more self-deprecatory, more self-preoccupied, and generally less content with themselves than *Ss* lower in the distribution of anxiety scores (Bendig, 1958; Cowen, Heilizer, Axelrod, & Sheldon, 1957; Doris & Sarason, 1955, Fiedler, et al., 1958; Holtzman & Bitterman, 1956; Holtzman, Calvin, & Bitterman, 1952; Trapp & Kausler, 1958; Westrope, 1953; Wolf, 1955). It may well be that highly motivating or ego-involving instructions serve the function of arousing these self-oriented tendencies. One recent study (Sarason, 1958a) has shown that *Ss* scoring high in test anxiety respond more positively to reassurance in an experimental situation than do low anxious *Ss*. A worthwhile problem for future research would seem to be the development of techniques for the *extinction* rather than the arousal of anxiety responses.

Consistent with the interpretation of anxiety measures as indicators of sensitivity to implied personal threat is the finding by several investigators that there are no differences among groups differing in scores on anxiety scales when tested under neutral and apparently nonthreatening conditions (Axelrod, Cowen, & Heilizer, 1956; Sarason, 1956a, 1957a, 1957b; Silverman & Blitz, 1956). Sarason, in a series of three experiments (1956a, 1957a, 1957b) involving the effects of anxiety and experimental stress on verbal learning, failed to find under pre-experimental neutral conditions significant differences in performance between groups which differed in anxiety, although varying performance was obtained under later conditions of personal threat. This suggests a sensitivity interpretation of anxiety similar to the one offered by Davidson, Andrews, and Ross (1956). Furthermore, evidence recently reported suggests the possibility that the more directly related the content of items on the anxiety scale is to the situation in which *Ss* are to perform, the more useful is the measure of anxiety in showing interactions between scores on the scale and differential motivating instructions (Raphelson, 1957; Sarason, 1958a, 1959a, 1959c; Sarason & Palola, 1960).

The results of studies on anxiety and stress have led to what might be called a habit interpretation of anxiety (Child, 1954; Davidson, et al., 1956; S. A. Mednick, 1957; Nicholson, 1958; Sampson & Bindra, 1954; Sarason, 1958b, 1959a). This interpretation, briefly put, states that *Ss* scoring high and low in anxiety differ in the response tendencies activated by personally threatening conditions. Whereas low scoring *Ss* may react to such conditions with increased effort and attention to the task at hand, high scoring *Ss* respond to threat with self-oriented, personalized responses. More information is needed to clarify the conditions, such as those in the family and schoolroom environments, which are associated with the development of heightened responsiveness to stress. The rapidly burgeoning interest in the measurement of anxiety in children can be helpful in this regard (Castaneda, McCandless, & Palermo, 1956; McCandless, Castaneda, & Palermo, 1956; Sarason, Davidson, Lighthall, & Waite, 1958; Waite, Sarason, Lighthall, & Davidson, 1958).

A neglected problem in the creation of experimental stress situations is that of the *E* as an agent in creating a threat to *S*. Even when quite explicit motivating instructions are administered to *S*, there remains the problem of the administration of these instructions. The problem of variance among *E*s in the manner with which instructions are communicated cannot be overemphasized. Systematic study is needed of the relationship between *E* variables such as sex and personality of *E* and anxiety aroused.

Anxiety and Task Variables

As has already been mentioned, the originators of the MAS considered it to be a measure of drive, *D*, and were primarily interested in relating it to the concept of the response hierarchy. In simple, one-response situations such as eyelid conditioning, it was predicted that high anxious *S*s would perform at higher levels than would low anxious *S*s. However, as the complexity (e.g., intralist similarity) of the task to be learned increased, a superiority of low to high anxious *S*s was expected.

A number of the studies conducted within this framework have supported these assumptions (Farber & Spence, 1953; Montague, 1953; Ramond, 1953; Spence, Farber, & McFann, 1956; Taylor, 1951; Taylor & Spence, 1952). For example, Montague (1953) compared high and low anxiety groups in ability to learn lists of nonsense syllables which differed in association value and intralist similarity. A significant interaction was obtained with low anxious *S*s superior to high anxious *S*s on the most complex or difficult task. On the least complex task, high were superior to low anxious *S*s. These findings, although subject to alternative interpretations, are in accord with Hullian expectations. Farber (1955) and Taylor (1956) have presented summaries and analyses of work on anxiety from a drive point of view.

Despite these positive findings, a review of the literature also reveals a number of other studies either contradictory to or not consistent with a drive interpretation of anxiety (Axelrod, et al., 1956; Bindra, Paterson & Strzelecki, 1955; Deese, Lazarus, & Keenan, 1953; Heilizer, Axelrod, & Cowen, 1956; Kamin & Clark, 1957; Kamin & Fedorchak, 1957; Saltz & Hoehn, 1957; Silverman & Blitz, 1956). Several of these studies were specifically designed to test predictions from Hullian theory concerning the performance on either simple or complex tasks of groups scoring high and low on anxiety scales. An especially interesting experiment was performed by Bindra, Paterson, and Strzelecki (1955). They did not obtain significant differences between high and low anxious *S*s in simple conditioning. It is interesting that their situation involved a nondefensive response rather than the defensive one used in many drive studies. The threatening aspects of receiving puffs of air in the region of the eye may be much more crucial in affecting performance than the lack of response hierarchy competition hypothesized for such one-response situations (Hilgard, Jones, & Kaplan, 1951; Kamin, 1955). It seems likely that for certain tasks there exists a confounding of task simplicity with task stressfulness. Korchin and Levine (1957) have actually interpreted complexity of the learning situation not so much as a task variable but as a stress variable. Kausler and Trapp (1959) have recently presented a critique of the drive interpretation of anxiety which discusses other problems along these lines.

One particular problem suggested

by studies of anxiety in which complex tasks are used is this dual aspect of task complexity. A complex task can be difficult *and* at least potentially threatening to *S*. Under what conditions either or both of these aspects of task complexity are operative has as yet not been systematically studied. Certainly a closer tie-in between studies of anxiety and stress and studies of anxiety and task factors seems indicated.

In an attempt in this direction, Sarason and Palola (1960) manipulated simultaneously the variables of anxiety, differential motivating instructions, and task complexity in three experiments. Significant triple interactions involving the three variables studied were obtained in every case. These results are in accord with the dual properties of task complexity already mentioned. They seem to suggest the necessity of developing an integrated interpretation of anxiety in terms of the experimental conditions most detrimental to the performance of high anxious *S*s. For example, the combined use of high threat and high complexity of task might lead to larger differences in performance between high and low anxious *S*s than the manipulation of either threat or complexity alone. A study by Taylor (1958) illustrates the need for this type of research, and Nicholson (1958) has recently presented findings consistent with this formulation.

In addition to needed advances in theory in integrating the anxiety, motivational, and task variables, it is imperative that theories of anxiety also incorporate such variables as the sex of *S* and *E*. This was suggested by Kamin and Clark (1957) and has been most dramatically illuminated by the results of a group of researchers at the University of Rochester (Axelrod, et al., 1956; Heilizer, et al., 1956). These workers have consistently shown significant interactions between (*a*) anxiety scores, (*b*) sex of *S*, and (*c*) *E* characteristics. These latter two variables related more powerfully to anxiety of *S*s than did task complexity, the primary focus of their research. As these authors point out, psychological theory has failed to deal systematically with the *S* and *E* variables.

Anxiety and Intelligence

Although several investigators have reported negative relationships between MAS scores and intellectual performance for certain *S* populations (Grice, 1955; Kerrick, 1955; Matarazzo, Ulett, Guze, & Saslow, 1954; Siegman, 1956a, 1956b; Spielberger, 1958), the majority of studies relating measures of general anxiety to measures of intellectual performance have yielded nonsignificant correlations (Dana, 1957; Davids & Eriksen, 1955; Goodstein & Farber, 1957; Jackson & Bloomberg, 1958; Klugh & Bendig, 1955; Matarazzo, 1955; Sarason, 1956b, 1959a; Schulz & Calvin, 1955; Taylor, 1955).

Whether one should infer that high anxious *S*s are less bright than other *S*s when significant negative correlations between anxiety and intellectual performance are obtained depends on the interpretation placed on anxiety scales. The finding that under stressful conditions low anxious *S*s perform at higher levels than high anxious *S*s, and under nonstressful conditions high and low anxious *S*s perform equally well, might suggest that labeling a test as an intelligence test and the difficulty of the test itself may arouse anxiety responses in high anxious *S*s which interfere with their performance. Motivational and situational variables associated with testing have not yet been manipulated systematically in studies

which attempt to relate anxiety and intelligence.

As was indicated earlier it would appear that, for college students, tests of the ACE type are unrelated to, or only very slightly related to, measures of general anxiety such as MAS. However, studies which have related test anxiety, i.e., anxiety experienced in test situations, to measures of intellectual performance have shown consistent negative correlations. The *Ss* scoring high in test anxiety obtain lower performance scores than *Ss* with lower scores (Cowen, 1957; Mandler & Cowen, 1958; Sarason, 1957c, 1959a; Sarason & Mandler, 1952). In one study (Sarason, 1959a) in which both general and test anxiety indices were used, it was found that test anxiety correlated negatively with several intellectual measures for both male and female college students, but measures of general anxiety and other personality variables were unrelated to intelligence.

An important problem in the study of the correlation between anxiety and intelligence which has not been given enough emphasis is that of the range of intellectual ability studied. If restricted ranges of ability are used, it will make it less likely that significant correlations will emerge. Although investigators in this field are aware of the limited inferences one can draw on the basis of restricted sampling (e.g., college students, air force recruits, student nurses), no systematic attempts have been made to study the relationships between anxiety and intelligence in different populations using similar measures of anxiety and intelligence in all comparisons. Spielberger (1958) and Calvin, Koons, Bingham, and Fink (1955) have presented evidence which strongly suggests the need for such a systematic consideration of sampling variations.

Anxiety and Physiological Variables

As anxiety is defined clinically, it is typically assumed that it has important physiological correlates. On the basis of assumptions of this type, several investigators have sought relationships between anxiety and a variety of physiological measures (e.g., GSR). Although work in this area seems only to be getting under way, the results to date have been largely negative. Measures of questionnaire-defined anxiety such as MAS do not seem to relate consistently to physiological responding (Beam, 1955; Berry & Martin, 1957; Calvin, McGuigan, Tyrrell, & Soyars, 1956; Lotsof & Downing, 1956; Raphelson, 1957). Although these negative findings can be taken as reflecting poorly on the validity of MAS-type scales, it may also be that these scales are tapping aspects of anxiety other than autonomic functioning. It is known that there are marked individual differences among *Ss* in their physiological response patterns under stress conditions (Lacey, 1950; Lacey, Bateman, & Van Lehn, 1953). Consequently, in research relating anxiety and autonomic response, it would seem desirable to study patterns of physiological responding rather than only one physiological response measure.

Another important variable as yet unstudied in this area relates to the conditions under which *Ss*' physiological responses are measured (Martin & McGowan, 1955). The situational and experimental conditions under which an hypothesized relationship should be present or not present have not been explored. It is known that even patients diagnosed

as anxiety states do not display anxiety symptoms at all times and do not always show the same patterns of symptoms. Just as the habit interpretation of anxiety would expect that, if one wished to maximize differences between high and low anxious *Ss* on intelligence tests, *Ss* would have to be run using highly motivating, ego-involving conditions, so also physiological differences between *Ss* differing in anxiety might occur only under stressful or motivating conditions.

Measurement of Anxiety and Its Relation to Personality and Test Taking Attitudes

In constructing tests of personality we must ask ourselves many questions related to their reliability and validity. What does the test purport to measure? What is the best format for the test? How does the test relate to other available instruments? What are the best ways in which to establish the validity of the test?

Jessor and Hammond (1957) have pointed out in relation to anxiety scales that some of these questions can ultimately be answered through the process of construct validation. Unfortunately, at present, the construct validation of anxiety scales is at a rudimentary stage. For example, is a true-false paper and pencil test the most appropriate measure of anxiety? At present we do not know the answer to this question. Probably the major reason for the wide use of paper and pencil indices of anxiety is convenience. While convenience is a desirable characteristic, research is needed to investigate less convenient but perhaps more useful indices.

Perhaps the most parsimonious statement that one can make concerning what is measured by existing scales of anxiety is that they measure the extent to which an individual is willing to admit to experiencing anxiety in certain situations. However, also to be considered are the following possibilities: (*a*) high anxiety scores may be obtained by certain *Ss* because of plus-getting tendencies, i.e., tendencies to attribute "bad" characteristics to themselves; (*b*) high scores may be obtained by particularly frank and open *Ss*; (*c*) high scores may be obtained by *Ss* who are particularly perceptive of their own reactions. The converse of each of these possibilities represents a possible basis for low anxiety scores.

In this connection, it should be pointed out that many true-false scales of anxiety have been found to correlate very highly and negatively with measures of defensiveness, test-taking attitude, and the tendency to respond to personality test items in a socially desirable direction (Edwards, 1957; Fordyce, 1956). Such high correlations may indicate that anxiety scores are explainable in terms of test-taking attitude. Whether or not this is true is a problem that construct validation studies should be designed to answer.

Interestingly, it is possible to construct scales of anxiety which do not correlate very highly with measures of test-taking attitude. The writer (Sarason, 1959*a*) has obtained correlations between the Test Anxiety Scale and *SD* of $-.49$ for women and $-.23$ for men. Also several forced-choice anxiety scales have been presented which to a very considerable extent seem to reduce the correlations between anxiety scores and measures of test-taking attitude (Christie & Budnitzky, 1957; Heineman, 1953; Lykken, 1957; Silverman, 1957). However, it is possible that forced-choice techniques in the field of

personality measurement create as many problems as they solve (Guilford, 1959, pp. 188–189). More research designed to measure anxiety in a variety of ways and to better understand anxiety and test-taking attitude relationships seems indicated.

Two additional areas which require further study are the relationship of measures of anxiety to (*a*) other personality dimensions and (*b*) to the clinical conditions of patients. With respect to anxiety and other personality measures, it appears that at least one test, the Psychasthenia, *Pt*, scale of the MMPI, correlates as highly with the MAS as the MAS correlates with itself (Brackbill & Little, 1954; Deese, et al., 1953; Eriksen & Davids, 1955). Although *Pt*-MAS item overlap is clearly a factor in these high correlations, this relationship between MAS and *Pt* may suggest that high scorers on anxiety scales obtain such scores because of ruminative, obsessive thinking about themselves. If scales of anxiety, or at least the MAS, are measuring a variable related to obsessive-compulsive tendencies, then the positive correlations reported by some investigators (Davids, 1955b; Siegal, 1954) between MAS and measures of authoritarianism perhaps might be explained in terms of the dogmatism and rigidity often observed in neurotic obsessive-compulsive personalities.

As was mentioned earlier, the weight of the available evidence indicates that scales of anxiety are tapping tendencies towards neuroticism, maladjustment, and self-dissatisfaction (Bendig, 1958; Cowen, et al., 1957; Davids & Eriksen, 1955; Fiedler, et al., 1958; Holtzman, et al., 1952; Winne, 1951). There are indications also that this heightened insecurity of high anxious individuals may result in a greater susceptibility to persuasion and opinion change, and to greater sensitivity to reinforcements provided by *E* to *S* in learning situations (Fine, 1957; Janis, 1955; Sarason, 1958b; Taffel, 1955). For example, in two similar verbal conditioning studies both Taffel (1955) and Sarason (1958b) found that high anxious neuropsychiatric patients changed their frequency of usage of a verbal response class reinforced by *E* more easily than did patients with lower anxiety scores.

These sorts of relationships are consistent with the observations made in psychotherapeutic contacts with patients that liklihood of movement in therapy is, to a considerable extent, positively related to the patient's anxiety level. However, results of studies on the diagnostic value of indices of anxiety do not as yet fall into clearly discernible patterns, and it is hard to draw generalizations concerning the value of these indices as diagnostic tools. It can be said that a number of investigators have found anxiety scales to be correlated either with indices of general maladjustment or ratings of anxiety made by clinicians (Buss, Weiner, Durkee, & Baer, 1955; Holtzman & Bitterman, 1956; Lauterbach, 1958; Matarazzo, Guze, & Matarazzo, 1955; Taft, 1957). The magnitude of these correlations, while significant, has often been so low as to preclude use in the individual case. Kendall (1954) has suggested that MAS be regarded as only a rough clinical tool.

An important methodological problem in relating anxiety indices to the ratings of patients' anxiety made by clinicians is the method of obtaining such judgments. Poorly constructed rating scales will inevitably lead to low-order relationships with other

measures. In this regard, attention should be called to the interesting study by Buss et al. (1955) in which the use of adequate procedures to insure interrater reliability among clinicians no doubt contributed to the positive results obtained.

Summary

This paper has dealt with the relationship of anxiety to certain research areas. Existing research suggests the following summaries:

1. The performance of high anxious *Ss* is detrimentally affected by verbally administered highly motivating communications. This result is consistent with the view that high anxious *Ss* emit personalized, self-oriented interfering responses when threat is perceived in the environment. Under nonthreat conditions the emission of such responses would not be expected. It was pointed out that several methodological problems remain to be solved in the assessment of the relationship between anxiety and stress. On the *E* side there is the confounding of variables such as experimental instructions with characteristics of the *E* administering such instructions. On the *S* side, more must be learned about the relationship of sex and personality characteristics of *Ss* which affect their responses to conditions of implied threat.

2. The results of several experiments using MAS as a measure of drive have indicated that, as task complexity increases, the disadvantage of high to low anxious *Ss* appears to increase. However, there has been considerable research in which this relationship was not confirmed. Perhaps the major theoretical problem in the anxiety-task complexity relationship is the interpretation to be placed on the complexity variable. Complex tasks can be both difficult and emotionally arousing. It would appear that both of these aspects of task complexity must be considered.

3. Although several reports of correlations between measures of general anxiety, such as the MAS, and intellectual measures are to be found in the literature, it does not appear that this relationship consistently holds. Specific test anxiety, on the other hand, does seem to relate negatively to intellectual measures. It has been suggested that indices of specific anxieties such as test anxiety may prove more valuable for specific purposes than more general indices like MAS.

4. Negative findings seem to pervade the study of the relationship of anxiety to physiological indices. The typical procedure has been to select *Ss* differing in anxiety scores and to compare these *Ss* on autonomic measures such as GSR. It was suggested that the lack of significant relationships in such comparisons may be attributable to a failure to make the comparisons under conditions of perceived threat or stress. High and low anxious *Ss* may differ in physiological response under threat but not under nonthreat conditions.

5. Problems of the effects of test-taking attitudes on anxiety scores and the format of anxiety scales have as yet not been given the intensive study which they merit. While most indices of anxiety of the MAS type have been found to correlate negatively and very highly with measures of test-taking attitudes (e.g., the *K* scale of the MMPI), this has not been obtained in all cases. Forced-choice techniques and the Test Anxiety Scale do not correlate as highly with test-taking attitude as do MAS and other general anxiety indices. Further construct validation of both

anxiety and test-taking attitude scales may illuminate the significance of these findings.

The aim of this paper has been to point to some of the consistencies and inconsistencies in the area of anxiety research and to suggest some of the uncontrolled and confounding variables which may have led to discrepant findings and which need to be systematically studied in future research.

REFERENCES

AXELROD, H. D., COWEN, E. L., & HEILIZER, F. The correlates of manifest anxiety in stylus maze learning. *J. exp. Psychol.*, 1956, **51**, 131–138.

BEAM, J. C. Serial learning and conditioning under real-life stress. *J. abnorm. soc. Psychol.*, 1955, **51**, 543–551.

BENDIG, A. W. The development of a short form of the Manifest Anxiety Scale. *J. consult. Psychol.*, 1956, **20**, 384.

BENDIG, A. W. Identification of item factor patterns within the Manifest Anxiety Scale. *J. consult. Psychol.*, 1958, **22**, 158.

BENDIG, A. W., & VAUGHAN, C. J. Manifest anxiety, discrimination, and transportation. *Amer. J. Psychol.*, 1957, **70**, 286–288.

BERRY, J. C., & MARTIN, B. GSR reactivity as a function of anxiety, instructions, and sex. *J. abnorm. soc. Psychol.*, 1957, **54**, 9–12.

BINDRA, D., PATERSON, A. L., & STRZELECKI, JOANNA. On the relation between anxiety and conditioning. *Canad. J. Psychol.*, 1955, **9**, 1–6.

BLAKE, R. R., & MOUTON, JANE S. Personality. In *Annu. Rev. Psychol.*, 1959, **10**, 203–232.

BRACKBILL, G., & LITTLE, K. B. MMPI correlates of the Taylor scale of manifest anxiety. *J. consult. Psychol.*, 1954, **18**, 433–436.

BUSS, A., WEINER, M., DURKEE, ANN, & BAER, M. The measurement of anxiety in clinical situations. *J. consult. Psychol.*, 1955, **19**, 125–129.

CALVIN, A. D., KOONS, P. B., JR., BINGHAM, J. L., & FINK, H. H. A further investigation of the relationship between manifest anxiety and intelligence. *J. consult. Psychol.*, 1955, **19**, 280–282.

CALVIN, A. D., MCGUIGAN, F. J., TYRRELL, SYBIL, & SOYARS, M. Manifest anxiety and the palmar perspiration index. *J. consult. Psychol.*, 1956, **20**, 356.

CASTANEDA, A., MCCANDLESS, B. R., & PALERMO, D. S. The children's form of the Manifest Anxiety Scale. *Child Develpm.*, 1956, **27**, 317–326.

CHILD, I. L. Personality. *Annu. Rev. Psychol.*, 1954, **5**, 149–170.

CHRISTIE, R., & BUDNITZKY, S. A short forced-choice anxiety scale. *J. consult. Psychol.*, 1957, **21**, 501.

COWEN, E. L., HEILIZER, F., AXELROD, H. S., & SHELDON, A. The correlates of manifest anxiety in perceptual reactivity, rigidity, and self-conflict. *J. consult. Psychol.*, 1957, **21**, 405–411.

COWEN, JUDITH E. Test anxiety in high school students and its relationship to performance on group tests. Unpublished doctoral dissertation, School of Education, Harvard Univer., 1957.

COX, F., & SARASON, S. B. Test anxiety and Rorschach performance. *J. abnorm. soc. Psychol.*, 1954, **49**, 371–377.

CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281–302.

DANA, R. H. Manifest anxiety, intelligence, and psychopathology. *J. consult. Psychol.*, 1957, **21**, 38–40.

DAVIDS, A. Relations among several objective measures of anxiety under different conditions of motivation. *J. consult. Psychol.*, 1955, **19**, 275–279. (a)

DAVIDS, A. Some personality and intellectual correlates of intolerance of ambiguity. *J. abnorm. soc. Psychol.*, 1955, **51**, 415–420. (b)

DAVIDS, A., & ERIKSEN, C. W. The relation of manifest anxiety to association productivity and intellectual attainment. *J. consult. Psychol.*, 1955, **19**, 219–222.

DAVIDSON, W. Z., ANDREWS, J. A., & ROSS, S Effects of stress and anxiety on continuous high speed color naming. *J. exp. Psychol.*, 1956, **52**, 13–17.

DEESE, J., LAZARUS, R. S., & KEENAN, J. Anxiety, anxiety reduction, and stress in learning. *J. exp. Psychol.*, 1953, **46**, 55–60.

DIXON, J. J., DEMONCHAUX, C., & SANDLER, J. Patterns of anxiety: An analysis of social anxieties. *Brit. J. med. Psychol.*, 1957, **30**, 107–112.

DORIS, J., & SARASON, S. B. Test anxiety and blame assignment in a failure situation. *J. abnorm. soc. Psychol.*, 1955, **50**, 335–338.

EDWARDS, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

EICHHORN, R., & TRACKTIR, J. The relationship between anxiety, hypnotically induced emotions and gastric secretion. *Gastroenterology*, 1955, **29**, 422–431.

ERIKSEN, C. W., & DAVIDS, A. The meaning and clinical validity of the Taylor anxiety scale and the hysteria-psychasthenia scales from the MMPI. *J. abnorm. soc. Psychol.*, 1955, **50**, 135–137.

ERIKSEN, C. W., & WECHSLER, H. Some effects of experimentally induced anxiety upon discrimination behavior. *J. abnorm. soc. Psychol.*, 1955, **51**, 458–463.

FARBER, I. E. The role of motivation in verbal learning and performance. *Psychol. Bull.*, 1955, **52**, 311–327.

FARBER, I. E., & SPENCE, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 120–125.

FARBER, I. E., & SPENCE, K. W. Effects of anxiety, stress, and task variables on reaction time. *J. Pers.*, 1956, **25**, 1–18.

FELDMAN, M. J., & SIEGEL, S. M. The effect on self-description of combining anxiety and hostility items on a single scale. *J. clin. Psychol.*, 1958, **14**, 74–77.

FIEDLER, F. E., DODGE, JOAN S., JONES, R. E., & HUTCHINS, E. B. Interrelations among measures of personality adjustment in nonclinical populations. *J. abnorm. soc. Psychol.*, 1958, **56**, 345–351.

FINE, B. J. Conclusion-drawing, communication, credibility, and anxiety as factors in opinion change. *J. abnorm. soc. Psychol.*, 1957, **54**, 369–374.

FORDYCE, W. E. Social desirability in the MMPI. *J. consult. Psychol.*, 1956, **20**, 171–175.

GOODSTEIN, L. D. Interrelationship among several measures of anxiety and hostility. *J. consult. Psychol.*, 1954, **18**, 35–39.

GOODSTEIN, L. D., & FARBER, I. E. On the relation between A-scale scores and digit symbol performance. *J. consult. Psychol.*, 1957, **21**, 152–154.

GORDON, E. M., & SARASON, S. B. The relationship between test anxiety and other anxieties. *J. Pers.*, 1955, **23**, 317–323.

GORDON, W. M., & BERLYNE, D. E. Drive level and flexibility in paired-associate nonsense-syllable learning. *Quart. J. exp. Psychol.*, 1954, **6**, 181–185.

GRICE, G. R. Discrimination reaction time as a function of anxiety and intelligence. *J. abnorm. soc. Psychol.*, 1955, **50**, 71–74.

GUILFORD, J. P. *Personality*. New York: McGraw Hill, 1959.

GYNTHER, RUTH A. The effects of anxiety and of situational stress on communicative efficiency. *J. abnorm. soc. Psychol.*, 1957, **54**, 274–276.

HEILIZER, F., AXELROD, H. S., & COWEN, E. L. The correlates of manifest anxiety in paired associate learning. *J. Pers.*, 1956, **24**, 463–474.

HEINEMAN, C. E. A forced-choice form of the Taylor anxiety scale. *J. consult. Psychol.*, 1953, **17**, 447–454.

HILGARD, E. R., JONES, L. V., & KAPLAN, S. J. Conditioned discrimination as related to anxiety. *J. exp. Psychol.*, 1951, **52**, 94–99.

HOLTZMAN, W. H., & BITTERMAN, M. E. A factorial study of adjustment to stress. *J. abnorm. soc. Psychol.*, 1956, **42**, 179–185.

HOLTZMAN, W. H., CALVIN, A. D., & BITTERMAN, M. E. New evidence for the validity of Taylor's Manifest Anxiety Scale. *J. abnorm. soc. Psychol.*, 1952, **47**, 853–854.

JACKSON, D. N., & BLOOMBERG, R. Anxiety: Unitas or multiplex? *J. consult. Psychol.*, 1958, **22**, 225.

JANIS, I. L. Anxiety indices related to susceptibility to persuasion. *J. abnorm. soc. Psychol.*, 1955, **51**, 663–667.

JENKINS, J. J., & LYKKEN, D. T. Individual differences. *Annu. Rev. Psychol.*, 1957, **8**, 79–112.

JENSEN, A. R. Personality. *Annu. Rev. Psychol.*, 1958, **9**, 295–322.

JESSOR, R., & HAMMOND, K. R. Construct validity and the Taylor anxiety scale. *Psychol. Bull.*, 1957, **54**, 161–170.

KAMIN, L. J. Relations between discrimination, apparatus stress, and the Taylor scale. *J. abnorm. soc. Psychol.*, 1955, **51**, 595–599.

KAMIN, L. J. The measurement of anxiety: A methodological note. *Canad. J. Psychol.*, 1957, **11**, 71–74.

KAMIN, L. J., & CLARK, J. W. The Taylor scale and reaction time. *J. abnorm. soc. Psychol.*, 1957, **54**, 262–263.

KAMIN, L. J., & FEDORCHAK, OLGA. The Taylor scale, hunger, and verbal learning. *Canad. J. Psychol.*, 1957, **11**, 212–218.

KAUSLER, D. H., & TRAPP, E. P. Methodological considerations in the construct validation of drive-oriented scales. *Psychol. Bull.* 1959, **56**, 152–157.

KENDALL, E. The validity of Taylor's Manifest Anxiety Scale. *J. consult. Psychol.*, 1954, **18**, 429–432.

KERRICK, JEAN S. Some correlates of the Taylor Manifest Anxiety Scale. *J. abnorm. soc. Psychol.*, 1955, **50**, 75–77.

KLUGH, H. E., & BENDIG, A. W. The manifest anxiety and ACE scales and college achievement. *J. consult. Psychol.*, 1955, **19**, 487.

KORCHIN, S. J., & LEVINE, S. Anxiety and

verbal learning. *J. abnorm. soc. Psychol.*, 1957, **54**, 234–240.

LACEY, J. I. Individual differences in somatic response patterns. *J. comp. physiol. Psychol.*, 1950, **43**, 338–350.

LACEY, J. I., BATEMAN, DOROTHY, & VAN LEHN, RUTH. Autonomic response specificity. *Psychosom. Med.*, 1953, **15**, 8–21.

LAUTERBACH, C. G. The Taylor A scale and clinical measures of anxiety. *J. consult. Psychol.*, 1958, **22**, 314.

LOTSOF, E. J., & DOWNING, W. L. Two measures of anxiety. *J. consult. Psychol.*, 1956, **20**, 170.

LUCAS, J. D. The interactive effects of anxiety, failure, and intraserial duplication. *Amer. J. Psychol.*, 1952, **65**, 59–66.

LYKKEN, D. T. A study of anxiety in the sociopathic personality. *J. abnorm. soc. Psychol.*, 1957, **55**, 6–10.

MCCANDLESS, B. R., CASTANEDA, A., & PALERMO, D. S. Anxiety in children and social status. *Child Develpm.*, 1956, **27**, 385–391.

MCCLELLAND, D. C. Methods of measuring human motivation. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. Princeton, N. J.: Van Nostrand, 1958. Pp. 7–42.

MANDLER, G., & COWEN, JUDITH E. Test anxiety questionnaires. *J. consult. Psychol.*, 1958, **22**, 228–229.

MANDLER, G., & SARASON, S. B. A study of anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 166–173.

MARTIN, B., & MCGOWAN, B. Some evidence on the validity of the Sarason Test Anxiety Scale. *J. consult. Psychol.*, 1955, **19**, 468.

MATARAZZO, J. D., GUZE, S. B., & MATARAZZO RUTH G. An approach to the validity of the Taylor anxiety scale: Scores of medical and psychiatric patients. *J. abnorm. soc. Psychol.*, 1955, **51**, 276–280.

MATARAZZO, J. D., ULETT, G. A., GUZE, S. B., & SASLOW, G. The relationship between anxiety level and several measures of intelligence. *J. consult. Psychol.*, 1954, **18**, 201–205.

MATARAZZO, RUTH G. The relationship of manifest anxiety to Wechsler-Bellevue subtest performance. *J. consult. Psychol.*, 1955, **19**, 218.

MEDNICK, MARTHA T. Mediated generalization and the incubation effect as a function of manifest anxiety. *J. abnorm. soc. Psychol.*, 1957, **55**, 315–321.

MEDNICK, S. A. Generalization as a function of manifest anxiety and adaptation to psychological experiments. *J. consult. Psychol.*, 1957, **21**, 491–494.

MONTAGUE, E. K. The role of anxiety in serial rote learning. *J. exp. Psychol.*, 1953, **45**, 91–95.

NICHOLSON, W. M. The influence of anxiety upon learning. *J. Pers.*, 1958, **26**, 303–319.

RAMOND, C. Anxiety and task as determiners of verbal performance. *J. exp. Psychol.*, 1953, **46**, 120–124.

RAPHELSON, A. C. The relationships among imaginative, direct verbal, and physiological measures of anxiety in an achievement situation. *J. abnorm. soc. Psychol.*, 1957, **54**, 13–18.

ROSENBAUM, G. Stimulus generalization as a function of clinical anxiety. *J. abnorm. soc. Psychol.*, 1956, **53**, 281–285.

SALTZ, E., & HOEHN, A. J. A test of the Taylor-Spence theory of anxiety. *J. abnorm. soc. Psychol.*, 1957, **54**, 114–117.

SAMPSON, H., & BINDRA, D. "Manifest" anxiety, neurotic anxiety, and the rate of conditioning. *J. abnorm. soc. Psychol.*, 1954, **49**, 256–259.

SARASON, I. G. Effect of anxiety, motivational instructions, and failure on serial learning. *J. exp. Psychol.*, 1956, **51**, 253–260. (a)

SARASON, I. G. The relationship of anxiety and "lack of defensiveness" to intellectual performance. *J. consult. Psychol.*, 1956, **20**, 220–222. (b)

SARASON, I. G. The effect of anxiety and two kinds of failure on serial learning. *J. Pers.*, 1957, **25**, 383–392. (a)

SARASON, I. G. Effect of anxiety and two kinds of motivating instructions on verbal learning. *J. abnorm. soc. Psychol.*, 1957, **54**, 166–171. (b)

SARASON, I. G. Test anxiety, general anxiety, and intellectual performance. *J. consult. Psychol.*, 1957, **21**, 485–490. (c)

SARASON, I. G. Effects on verbal learning of anxiety, reassurance, and meaningfulness of material. *J. exp. Psychol.*, 1958, **56**, 472–477. (a)

SARASON, I. G. Interrelationships among individual difference variables, behavior in psychotherapy, and verbal conditioning. *J. abnorm. soc. Psychol.*, 1958, **56**, 339–344. (b)

SARASON, I. G. Intellectual and personality correlates of test anxiety. *J. abnorm. soc. Psychol.*, 1959, **59**, 272–275. (a)

SARASON, I. G. Relationships of measures of anxiety and experimental instructions to word association test performance. *J. abnorm. soc. Psychol.*, 1959, **59**, 37–42. (b)

SARASON, I. G. Relationships of measures of anxiety and experimental instructions to word association test performance: Further

findings. *Amer. Psychologist*, 1959, **14**, 370. (Abstract) (c)

SARASON, I. G., & PALOLA, E. G. The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. *J. exp. Psychol.*, 1960, **59**, 185–191.

SARASON, S. B., DAVIDSON, K. S., LIGHTHALL, F. F., & WAITE, R. R. A test anxiety scale for children. *Child Develpm.*, 1958, **29**, 105–113.

SARASON, S. B., & MANDLER, G. Some correlates of test anxiety. *J. abnorm. soc. Psychol.*, 1952, **47**, 810–817.

SARASON, S. B., MANDLER, G., & CRAIGHILL, P. C. The effect of differential instructions on anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 561–565.

SCHULZ, R. E., & CALVIN, A. D. A failure to replicate the findings of a negative correlation between manifest anxiety and ACE scores. *J. consult. Psychol.*, 1955, **19**, 223–224.

SIEGAL, S. Certain determinants and correlates of authoritarianism. *Genetic psychol. Monogr.*, 1954, **49**, 187–229.

SIEGMAN, A. W. Cognitive, affective, and psychopathological correlates of the Taylor MAS. *J. consult. Psychol.*, 1956, **20**, 137–141. (a)

SIEGMAN, A. W. The effect of manifest anxiety on a concept formation task, a non-directed learning task, and on timed and untimed intelligence tests. *J. consult. Psychol.*, 1956, **20**, 176–178. (b)

SILVERMAN, R. E. The Manifest Anxiety Scale as a measure of drive. *J. abnorm. soc. Psychol.*, 1957, **55**, 94–97.

SILVERMAN, R. E., & BLITZ, B. Learning and two kinds of anxiety. *J. abnorm. soc. Psychol.*, 1956, **52**, 301–303.

SINICK, D. Two anxiety scales correlated and examined for sex differences. *J. clin. Psychol.*, 1956, **12**, 394–395.

SPENCE, K. W., FARBER, I. E., & MCFANN, H. The relation of anxiety (drive) level to performance in competitional and non-competitional paired associates learning. *J. exp. Psychol.*, 1956, **52**, 296–305.

SPIELBERGER, C. D. On the relationship between manifest anxiety and intelligence. *J. consult. Psychol.*, 1958, **22**, 220–224.

TAFFEL, C. Anxiety and the conditioning of verbal behavior. *J. abnorm. soc. Psychol.*, 1955, **51**, 496–501.

TAFT, R. The validity of the Barron ego-strength scale and the Welsh anxiety index. *J. consult. Psychol.*, 1957, **21**, 247–249.

TAYLOR, JANET A. The relationship of anxiety to the conditioned eyelid response. *J. exp. Psychol.*, 1951, **41**, 81–92.

TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

TAYLOR, JANET A. The Taylor Manifest Anxiety Scale and intelligence. *J. abnorm. soc. Psychol.*, 1955, **51**, 347.

TAYLOR, JANET A. Drive theory and manifest anxiety. *Psychol. Bull.*, 1956, **53**, 303–320.

TAYLOR, JANET A. The effects of anxiety level and psychological stress on verbal learning. *J. abnorm. soc. Psychol.*, 1958, **57**, 55–60.

TAYLOR, JANET A., & SPENCE, K. W. The relationship of anxiety level to performance in serial learning. *J. exp. Psychol.*, 1952, **44**, 61–64.

TRAPP, E. P., & KAUSLER, P. H. Test anxiety level and goal-setting behavior. *J. consult. Psychol.*, 1958, **22**, 31–34.

TRUAX, C. B., & MARTIN, B. The immediate and delayed effect of failure as a function of task complexity and personalization of failure. *J. abnorm. soc. Psychol.*, 1957, **55**, 16–20.

WAITE, R. R., SARASON, S. B., LIGHTHALL, F. F., & DAVIDSON, K. S. A study of anxiety and learning in children. *J. abnorm. soc. Psychol.*, 1958, **57**, 267–270.

WELSH, G. S. An anxiety index and an internalization ratio for the MMPI. *J. consult. Psychol.*, 1952, **16**, 65–72.

WELSH, G. S. Factor dimensions *A* and *R*. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: Univer. Minnesota Press, 1956.

WESTROPE, MARTHA. Relations among Rorschach indices, manifest anxiety, and performance under stress. *J. abnorm. soc. Psychol.*, 1953, **48**, 515–524.

WINDLE, C. The relationships among five MMPI "anxiety" indices. *J. consult. Psychol.*, 1955, **19**, 61–63.

WINNE, J. F. A scale of neuroticism: An adaptation of the Minnesota Multiphasic Personality Inventory. *J. clin. Psychol.*, 1951, **7**, 117–122.

WOLF, W. M. Certainty: Generality and relation to manifest anxiety. *J. abnorm. soc. Psychol.*, 1955, **50**, 59–64.

ZIMET, C. N., & BRACKBILL, G. A. The role of anxiety in psychodiagnosis. *J. clin. Psychol.*, 1956, **12**, 173–177.

(Received June 17, 1959)

Psychological Bulletin
1960, Vol. 57, No. 5, 416–428

THE FALLACY OF THE NULL-HYPOTHESIS SIGNIFICANCE TEST

WILLIAM W. ROZEBOOM
St. Olaf College

The theory of probability and statistical inference is various things to various people. To the mathematician, it is an intricate formal calculus, to be explored and developed with little professional concern for any empirical significance that might attach to the terms and propositions involved. To the philosopher, it is an embarrassing mystery whose justification and conceptual clarification have remained stubbornly refractory to philosophical insight. (A famous philosophical epigram has it that induction [a special case of statistical inference] is the glory of science and the scandal of philosophy.) To the experimental scientist, however, statistical inference is a research instrument, a processing device by which unwieldy masses of raw data may be refined into a product more suitable for assimilation into the corpus of science, and in this lies both strength and weakness. It is strength in that, as an ultimate *consumer* of statistical methods, the experimentalist is in position to demand that the techniques made available to him conform to his actual needs. But it is also weakness in that, in his need for the tools constructed by a highly technical formal discipline, the experimentalist, who has specialized along other lines, seldom feels competent to extend criticisms or even comments; he is much more likely to make unquestioning application of procedures learned more or less by rote from persons assumed to be more knowledgeable of statistics than he. There is, of course, nothing surprising or reprehensible about this—one need not understand the principles of a complicated tool in order to make effective use of it, and the research scientist can no more be expected to have sophistication in the theory of statistical inference than he can be held responsible for the principles of the computers, signal generators, timers, and other complex modern instruments to which he may have recourse during an experiment. Nonetheless, this leaves him particularly vulnerable to misinterpretation of his aims by those who build his instruments, not to mention the ever present dangers of selecting an inappropriate or outmoded tool for the job at hand, misusing the proper tool, or improvising a tool of unknown adequacy to meet a problem not conforming to the simple theoretical situations in terms of which existent instruments have been analyzed. Further, since behaviors once exercised tend to crystallize into habits and eventually traditions, it should come as no surprise to find that the tribal rituals for data-processing passed along in graduate courses in experimental method should contain elements justified more by custom than by reason.

In this paper, I wish to examine a dogma of inferential procedure which, for psychologists at least, has attained the status of a religious conviction. The dogma to be scrutinized is the "null-hypothesis significance test" orthodoxy that passing statistical judgment on a scientific hypothesis by means of experimental observa-

tion is a decision procedure wherein one rejects or accepts a null hypothesis according to whether or not the value of a sample statistic yielded by an experiment falls within a certain predetermined "rejection region" of its possible values. The thesis to be advanced is that despite the awesome pre-eminence this method has attained in our experimental journals and textbooks of applied statistics, it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research. This is not a particularly original view—traditional null-hypothesis procedure has already been superceded in modern statistical theory by a variety of more satisfactory inferential techniques. But the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene.

To examine the method in question in greater detail, and expose some of the discomfitures to which it gives rise, let us begin with a hypothetical case study.

A Case Study in Null-Hypothesis Procedure; or, A Quorum of Embarrassments

Suppose that according to the theory of behavior, T_0, held by most right-minded, respectable behaviorists, the extent to which a certain behavioral manipulation M facilitates learning in a certain complex learning situation C should be null. That is, if "ϕ" designates the degree to which manipulation M facilitates the acquisition of habit H under circumstances C, it follows from the orthodox theory T_0 that $\phi = 0$. Also suppose, however, that a few radicals have persistently advocated an alternative theory T_1 which entails, among other things, that the facilitation of H by M in circumstances C should be appreciably greater than zero, the precise extent being dependent upon the values of certain parameters in C. Finally, suppose that Igor Hopewell, graduate student in psychology, has staked his dissertation hopes on an experimental test of T_0 against T_1 on the basis of their differential predictions about the value of ϕ.

Now, if Hopewell is to carry out his assessment of the comparative merits of T_0 and T_1 in this way, there is nothing for him to do but submit a number of *S*s to manipulation M under circumstances C and compare their efficiency at acquiring habit H with that of comparable *S*s who, under circumstances C, have *not* been exposed to manipulation M. The difference, d, between experimental and control *S*s in average learning efficiency may then be taken as an operational measure of the degree, ϕ, to which M influences acquisition of H in circumstances C. Unfortunately, however, as any experienced researcher knows to his sorrow, the interpretation of such an observed statistic is not quite so simple as that. For the observed dependent variable d, which is actually a performance measure, is a function not only of the extent to which M influences acquisition of H, but of many additional major and minor factors as well. Some of these, such as deprivations, species, age, laboratory conditions, etc., can be removed from consideration by holding them essentially constant. Others, however, are not so easily controlled, especially those customarily subsumed under the headings of "individual differences" and "errors of measurement." To

curtail a long mathematical story, it turns out that with suitable (possibly justified) assumptions about the distributions of values for these uncontrolled variables, the manner in which they influence the dependent variable, and the way in which experimental and control *Ss* were selected and manipulated, the observed sample statistic d may be regarded as the value of a normally distributed random variate whose average value is ϕ and whose variance, which is independent of ϕ, is unbiasedly estimated by the square of another sample statistic, s, computed from the data of the experiment.[1]

The import of these statistical considerations for Hopewell's dissertation, of course, is that he will not be permitted to reason in any simple way from the observed d to a conclusion about the comparative merits of of T_0 and T_1. To conclude that T_0, rather than T_1, is correct, he must argue that $\phi = 0$, rather than $\phi > 0$. But the observed d, whatever its value, is logically compatible both with the hypothesis that $\phi = 0$ and the hypothesis that $\phi > 0$. How then, can Hopewell use his data to make a comparison of T_0 and T_1? As a well-trained student, what he *does*, of course, is to divide d by s to obtain what, under H_0, is a t statistic, consult a table of the t distributions under the appropriate degrees-of-freedom, and announce his experiment as disconfirming or supporting T_0, respectively, according to whether or not the discrepancy between d and the zero value expected under T_0 is "statistically significant"—i.e., whether or not the observed value of d/s falls outside of the interval between two extreme percentiles (usually the 2.5th and 97.5th) of the t distribution with that df. If asked by his dissertation committee to justify this behavior, Hopewell would rationalize something like the following (the more honest reply, that this is what he has been taught to do, not being considered appropriate to such occasions):

> In deciding whether or not T_0 is correct, I can make two types of mistakes: I can reject T_0 when it is in fact correct [Type I error], or I can accept T_0 when in fact it is false [Type II error]. As a scientist, I have a professional obligation to be cautious, but a 5% chance of error is not unduly risky. Now if all my statistical background assumptions are correct, then, if it is really true that $\phi = 0$ as T_0 says, there is only one chance in 20 that my observed statistic d/s will be smaller than $t_{.025}$ or larger than $t_{.975}$, where by the latter I mean, respectively, the 2.5th and 97.5th percentiles of the t distribution with the same degrees-of-freedom as in my experiment. Therefore, if I reject T_0 when d/s is smaller than $t_{.025}$ or larger than $t_{.975}$, and accept T_0 otherwise, there is only a 5% chance that I will reject T_0 incorrectly.

If asked about his Type II error, and why he did not choose some other rejection region, say between $t_{.475}$ and $t_{.825}$, which would yield the same probability of Type I error, Hopewell should reply that although he has no way to compute his probability of Type II error under the assumptions traditionally authorized by null-hypothesis procedure, it is presumably minimized by taking the rejection region at the extremes of the t distribution.

Let us suppose that for Hopewell's data, $d = 8.50$, $s = 5.00$, and $df = 20$. Then $t_{.975} = 2.09$ and the acceptance region for the null hypothesis $\phi = 0$ is $-2.09 < d/s < 2.09$, or $-10.45 < d < 10.45$. Since d does fall within this region, standard null-hypothesis decision procedure, which I shall henceforth abbreviate "NHD," dictates that the experiment is to be reported

[1] s is here the estimate of the standard error of the difference in means, not the estimate of the individual *SD*.

as supporting theory T_0. (Although many persons would like to conceive NHD testing to authorize only rejection of the hypothesis, not, in addition, its acceptance when the test statistic fails to fall in the rejection region, if failure to reject were not taken as grounds for acceptance, then NHD procedure would involve no Type II error, and no justification would be given for taking the rejection region at the extremes of the distribution, rather than in its middle.) But even as Hopewell reaffirms T_0 in his dissertation, he begins to feel uneasy. In fact, several disquieting thoughts occur to him:

1. Although his test statistic falls within the orthodox acceptance region, a value this divergent from the expected zero should nonetheless be encountered less than once in 10. To argue in favor of a hypothesis on the basis of data ascribed a p value no greater than .10 (i.e., 10%) by that hypothesis certainly does not seem to be one of the more impressive displays of scientific caution.

2. After some belated reflection on the details of theory T_1, Hopewell observes that T_1 not only predicts that $\phi > 0$, but with a few simplifying assumptions no more questionable than is par for this sort of course, the value that ϕ should have can actually be computed. Suppose the value derived from T_1 in this way is $\phi = 10.0$. Then, rather than taking $\phi = 0$ as the null hypothesis, one might just as well take $\phi = 10.0$; for under the latter, $(d - 10.0)/s$ is a 20 *df t* statistic, giving a two-tailed, 95% significance, acceptance region for $(d - 10.0)/s$ between $-.209$ and 2.09. That is, if one lets T_1 provide the null hypothesis, it is accepted or rejected according to whether or not $-.45 < d < 20.45$, and by this latter test, therefore, Hopewell's data must be taken to support T_1—in fact, the likelihood under T_1 of obtaining a test statistic this divergent from the expected 10.0 is a most satisfactory three chances in four. Thus it occurs to Hopewell that had he chosen to cast his professional lot with the T_1-ists by selecting $\phi = 10.0$ as his null hypothesis, he could have made a strong argument in favor of T_1 by precisely the same line of statistical reasoning he has used to support T_0 under $\phi = 0$ as the null hypothesis. That is, he could have made an argument that persons partial to T_1 would regard as strong. For behaviorists who are already convinced that T_0 is correct would howl that since T_0 is the dominant theory, only $\phi = 0$ is a legitimate null hypothesis. (And is it not strange that what constitutes a valid statistical argument should be dependent upon the majority opinion about behavior theory?)

3. According to the NHD test of a hypothesis, only two possible final outcomes of the experiment are recognized—either the hypothesis is rejected or it is accepted. In Hopewell's experiment, all possible values of d/s between -2.09 and 2.09 have the same interpretive significance, namely, indicating that $\phi = 0$, while conversely, all possible values of d/s greater than 2.09 are equally taken to signify that $\phi \neq 0$. But Hopewell finds this disturbing, for of the various possible values that d/s might have had, the significance of $d/s = 1.70$ for the comparative merits of T_0 and T_1 should surely be more similar to that of, say, $d/s = 2.10$ than to that of, say, $d/s = -1.70$.

4. In somewhat similar vein, it also occurs to Hopewell that had he opted for a somewhat riskier confidence level, say a Type I error of 10% rather than 5%, d/s would have fallen outside the region of accept-

ance and T_0 would have been rejected. Now surely the degree to which a datum corroborates or impugns a proposition should be independent of the datum-assessor's personal temerity. Yet according to orthodox significance-test procedure, whether or not a given experimental outcome supports or disconfirms the hypothesis in question depends crucially upon the assessor's tolerance for Type I risk.

Despite his inexperience, Igor Hopewell is a sound experimentalist at heart, and the more he reflects on these statistics, the more dissatisfied with his conclusions he becomes. So while the exigencies of graduate circumstances and publication requirements urge that his dissertation be written as a confirmation of T_0, he nonetheless resolves to keep an open mind on the issue, even carrying out further research if opportunity permits. And reading his experimental report, so of course would we—has any responsible scientist ever made up his mind about such a matter on the basis of a single experiment? Yet in this obvious way we reveal how little our actual inferential behavior corresponds to the statistical procedure to which we pay lip-service. For if we did, in fact, accept or reject the null hypothesis according to whether the sample statistic falls in the acceptance or in the rejection region, then there would be no replications of experimental designs, no multiplicity of experimental approaches to an important hypothesis—a single experiment would, by definition of the method, make up our mind about the hypothesis in question. And the fact that in actual practice, a single finding seldom even tempts us to such closure of judgment reveals how little the conventional model of hypothesis testing fits our actual evaluative behavior.

Decisions vs. Degrees of Belief

By now, it should be obvious that something is radically amiss with the traditional NHD assessment of an experiment's theoretical import. Actually, one does not have to look far in order to find the trouble—it is simply a basic misconception about the purpose of a scientific experiment. The null-hypothesis significance test treats acceptance or rejection of a hypothesis as though these were *decisions* one makes on the basis of the experimental data—i.e., that we elect to adopt one belief, rather than another, as a result of an experimental outcome. *But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested.* And even if the purpose of the experiment *were* to reach a decision, it could not be a decision to accept or reject the hypothesis, for decisions are voluntary commitments to action—i.e., are *motor* sets—whereas acceptance or rejection of a hypothesis is a *cognitive* state which may provide the basis for rational decisions, but is not itself arrived at by such a decision (except perhaps indirectly in that a decision may initiate further experiences which influence the belief).

The situation, in other words, is as follows: As scientists, it is our professional obligation to reason from available data to explanations and generalities—i.e., beliefs—which are supported by these data. But belief in (i.e., acceptance of) a proposition is not an all-or-none affair; rather, it is a matter of degree, and the extent to which a person believes or accepts a

proposition translates pragmatically into the extent to which he is willing to commit himself to the behavioral adjustments prescribed for him by the meaning of that proposition. For example, if that inveterate gambler, Unfortunate Q. Smith, has complete confidence that War Biscuit will win the fifth race at Belmont, he will be willing to accept any odds to place a bet on War Biscuit to win; for if he is absolutely *certain* that War Biscuit will win, then odds are irrelevant—it is simply a matter of arranging to collect some winnings after the race. On the other hand, the more that Smith has doubts about War Biscuit's prospects, the higher the odds he will demand before betting. That is, the *extent* to which Smith accepts or rejects the hypothesis that War Biscu t will win the fifth at Belmont is an important determinant of his betting decisions for that race.

Now, although a scientist's data supply *evidence* for the conclusions he draws from them, only in the unlikely case where the conclusions are logically deducible from or logically incompatible with the data do the data warrant that the conclusions be entirely accepted or rejected. Thus, e.g., the fact that War Biscuit has won all 16 of his previous starts is strong evidence in favor of his winning the fifth at Belmont, but by no means warrants the unreserved acceptance of this hypothesis. More generally, the data available confer upon the conclusions a certain *appropriate degree of belief*, and it is the inferential task of the scientist to pass from the data of his experiment to whatever *extent* of belief these and other available information justify in the hypothesis under investigation. In particular, the proper inferential procedure is *not* (except in the deductive case) a matter of deciding to accept (without qualification) or reject (without qualification) the hypothesis: even if adoption of a belief were a matter of voluntary action—which it is not—neither such extremes of belief or disbelief are appropriate to the data at hand. As an example of the disastrous consequences of an inferential procedure which yields only two judgment values, acceptance and rejection, consider how sad the plight of Smith would be if, whenever weighing the prospects for a given race, he always worked himself into either supreme confidence or utter disbelief that a certain horse will win. Smith would rapidly impoverish himself by accepting excessively low odds on horses he is certain will win, and failing to accept highly favorable odds on horses he is sure will lose. In fact, Smith's two judgment values need not be *extreme* acceptance and rejection in order for his inferential procedure to be maladaptive. All that is required is that the degree of belief arrived at be in general inappropriate to the likelihood conferred on the hypothesis by the data.

Now, the notion of "degree of belief appropriate to the data at hand" has an unpleasantly vague, subjective feel about it which makes it unpalatable for inclusion in a formalized theory of inference. Fortunately, a little reflection about this phrase reveals it to be intimately connected with another concept relating conclusion to evidence which, though likewise in serious need of conceptual clarification, has the virtues both of intellectual respectability and statistical familiarity. I refer, of course, to the *likelihood*, or *probability*, conferred upon a hypothesis by available evidence. Why should not Smith *feel*

certain, in view of the data available, that War Biscuit will win the fifth at Belmont? Because it *is* not certain that War Biscuit will win. More generally, what determines how strongly we should accept or reject a proposition is the probability given to this hypothesis by the information at hand. For while our voluntary actions (i.e., decisions) are determined by our intensities of belief in the relevant propositions, not by their actual probabilities, expected utility is maximized when the cognitive weights given to potential but not yet known-for-certain pay-off events are represented in the decision procedure by the probabilities of these events. We may thus relinquish the concept of "appropriate degree of belief" in favor of "probability of the hypothesis," and our earlier contention about the nature of data-processing may be rephrased to say that the proper inferential task of the experimental scientist is not a simple acceptance or rejection of the tested hypothesis, but determination of the probability conferred upon it by the experimental outcome. This likelihood of the hypothesis relative to whatever data are available at the moment will be an important determinant for decisions which must currently be made, but is not itself such a decision and is entirely subject to revision in the light of additional information.

In brief, what is being argued is that the scientist, whose task is not to prescribe actions but to establish rational beliefs upon which to base them, is fundamentally and inescapably committed to an explicit concern with the problem of inverse probability. What he wants to know is how plausible are his hypotheses, and he is interested in the probability ascribed by a hypothesis to an observed experimental outcome only to the extent he is able to reason backwards to the likelihood of the hypothesis, given this outcome. Put crudely, no matter how improbable an observation may be under the hypothesis (and when there are an infinite number of possible outcomes, the probability of any particular one of these is, usually, infinitely small—the familiar p value for an observed statistic under a hypothesis H is not actually the probability of that outcome under H, but a partial integral of the probability-density function of possible outcomes under H), it is still confirmatory (or at least nondisconfirmatory, if one argues from the data to rejection of the background assumptions) so long as the likelihood of the observation is even smaller under the alternative hypotheses. To be sure, the theory of hypothesis-likelihood and inverse probability is as yet far from the level of development at which it can furnish the research scientist with inferential tools he can apply mechanically to obtain a definite likelihood estimate. But to the extent a statistical method does not at least move in the *direction* of computing the probability of the hypothesis, given the observation, that method is not truly a method of *inference*, and is unsuited for the scientist's cognitive ends.

The Methodological Status of the Null-Hypothesis Significance Test

The preceding arguments have, in one form or another, raised several doubts about the appropriateness of conventional significance-test decision procedure for the aims it is supposed to achieve. It is now time to bring these charges together in an explicit bill of indictment.

1. The null-hypothesis significance

test treats "acceptance" or "rejection" of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a *degree* of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true.

2. It might be argued that the NHD test may nonetheless be regarded as a legitimate decision procedure if we translate "acceptance (rejection) of the hypothesis" as meaning "acting as though the hypothesis were true (false)." And to be sure, there are many occasions on which one must base a course of action on the credibility of a scientific hypothesis. (Should these data be published? Should I devote my research resources to and become identified professionally with this theory? Can we test this new Z bomb without exterminating all life on earth?) But such a move to salvage the traditional procedure only raises two further objections. (*a*) While the scientist—i.e., the person—must indeed make decisions, his *science* is a systematized body of (probable) *knowledge*, not an accumulation of decisions. The end product of a scientific investigation is a degree of confidence in some set of propositions, which then constitutes a *basis* for decisions. (*b*) Decision theory shows the NHD test to be woefully inadequate as a decision procedure. In order to decide most effectively when or when not to act as though a hypothesis is correct, one must know both the probability of the hypothesis under the data available and the utilities of the various decision outcomes (i.e., the values of accepting the hypothesis when it is true, of accepting it when it is false, of rejecting it when it is true, and of rejecting it when it is false). But traditional NHD procedure pays no attention to utilities at all, and considers the probability of the hypothesis, given the data—i.e., the inverse probability—only in the most rudimentary way (by taking the rejection region at the extremes of the distribution rather than in its middle). Failure of the traditional significance test to deal with inverse probabilities invalidates it not only as a method of rational inference, but *also* as a useful decision procedure.

3. The traditional NHD test unrealistically limits the significance of an experimental outcome to a mere two alternatives, confirmation or disconfirmation of the null hypothesis. Moreover, the transition from confirmation to disconfirmation as a function of the data is discontinuous—an arbitrarily small difference in the value of the test statistic can change its significance from confirmatory to disconfirmatory. Finally, the point at which this transition occurs is entirely gratuitous. There is absolutely no reason (at least provided by the method) why the point of statistical "significance" should be set at the 95% level, rather than, say the 94% or 96% level. Nor does the fact that we sometimes select a 99% level of significance, rather than the usual 95% level, mitigate this objection—one is as arbitrary as the other.

4. The null-hypothesis significance test introduces a strong bias in favor of one out of what may be a large number of reasonable alternatives. When sampling a distribution of unknown mean μ, different assumptions about the value of μ furnish an infi-

nite number of alternate null hypotheses by which we might assess the sample mean, and whichever hypothesis is selected is thereby given an enormous, in some cases almost insurmountable, advantage over its competitors. That is, NHD procedure involves an inferential double standard—the favored hypothesis is held innocent unless proved guilty, while any alternative is held guilty until no choice remains but to judge it innocent. What is objectionable here is not that some hypotheses are held more resistant to experimental extinction than others, but that the differential weighing is an all-or-none side effect of a personal choice, and especially, that the method *necessitates* one hypothesis being favored over all the others. In the classical theory of inverse probability, on the other hand, all hypotheses are treated on a par, each receiving a weight (i.e., its "a priori" probability) which reflects the credibility of that hypothesis on grounds other than the data being assessed.

5. Finally, if anything can reveal the practical irrelevance of the conventional significance test, it should be its failure to see genuine application to the inferential behavior of the research scientist. Who has ever given up a hypothesis just because one experiment yielded a test statistic in the rejection region? And what scientist in his right mind would ever feel there to be an appreciable difference between the interpretive significance of data, say, for which one-tailed $p = .04$ and that of data for which $p = .06$, even though the point of "significance" has been set at $p = .05$? In fact, the reader may well feel undisturbed by the charges raised here against traditional NHD procedure precisely because, without perhaps realizing it, he has never taken the method seriously anyway. Paradoxically, it is often the most firmly institutionalized tenet of faith that is most susceptible to untroubled disregard—in our culture, one must early learn to live with sacrosanct verbal formulas whose import for practical behavior is seldom heeded. I suspect that the primary reasons why null-hypothesis significance testing has attained its current ritualistic status are (*a*) the surcease of methodological insecurity afforded by having an inferential algorithm on the books, and (*b*) the fact that a by-product of the algorithm is so useful, and its end product so obviously inappropriate, that the latter can be ignored without even noticing that this has, in fact, been done. What has given the traditional method its spurious feel of usefulness is that the *first*, and by far most laborious, step in the procedure, namely, estimating the probability of the experimental outcome under the assumption that a certain hypothesis is correct, is also a crucial first step toward what one is genuinely concerned with, namely, an idea of the likelihood of that hypothesis, given this experimental outcome. Having obtained this most valuable statistical information under pretext of carrying through a conventional significance test, it is then tempting, though of course quite inappropriate, to heap honor and gratitude upon the method while overlooking that its actual *result*, namely, a decision to accept or reject, is not used at all.

Toward a More Realistic Appraisal of Experimental Data

So far, my arguments have tended to be aggressively critical—one can hardly avoid polemics when butchering sacred cows. But my purpose is

not just to be contentious, but to help clear the way for more realistic techniques of data assessment, and the time has now arrived for some constructive suggestions. Little of what follows pretends to any originality; I merely urge that ongoing developments along these lines should receive maximal encouragement.

For the statistical theoretician, the following problems would seem to be eminently worthy of research:

1. Of supreme importance for the theory of probability is analysis of what we mean by a proposition's "probability," relative to the evidence provided. Most serious students of the philosophical foundations of probability and statistics agree (cf. Braithwaite, pp. 119f.) that the probability of a proposition (e.g., the probability that the General Theory of Relativity is correct) does not, prima facie, seem to be the same sort of thing as the probability of an event-class (e.g., the probability of getting a head when this coin is tossed). Do the statistical concepts and formulas which have been developed for probabilities of the latter kind also apply to hypothesis likelihoods? In particular, are the probabilities of hypotheses quantifiable at all, and for the theory of inverse probability, do Bayes' theorem and its probability-density refinements apply to hypothesis probabilities? These and similar questions are urgently in need of clarification.

2. If we are willing to assume that Bayes' theorem, or something like it, holds for hypothesis probabilities, there is much that can be done to develop the classical theory of inverse probability. While computation of inverse probabilities turns essentially upon the parametric a priori probability function, which states the probability of each alternative hypothesis in the set under consideration prior to the outcome of the experiment, it should be possible to develop theorems which are invariant over important subclasses of a priori probability functions. In particular, the difference between the a priori probability function and the "a posteriori" probability function (i.e., the probabilities of the alternative hypotheses after the experiment), perhaps analyzed as a difference in "information," should be a potentially fruitful source of concepts with which to explore such matters as the "power" or "efficiency" of various statistics, the acquisition of inductive knowledge through repeated experimentation, etc. Another problem which seems to me to have considerable import, though not one about which I am sanguine, is whether inverse-probability theory can significantly be extended to hypothesis-probabilities, given knowledge which is only probabilistic. That is, can a theory of sentences of form "The probability of hypothesis H, given that E is the case, is p," be generalized to a theory of sentences of form "The probability of hypothesis H, given that the probability of E is q, is p"? Such a theory would seem to be necessary, e.g., if we are to cope adequately with the uncertainty attached to the background assumptions which always accompany a statistical analysis.

My suggestions for applied statistical analysis turn on the fact that while what is desired is the a posteriori probabilities of the various alternative hypotheses under consideration, computation of these by classical theory necessitates the corresponding a priori probability distribution, and in the more immediate future, at least, information about this will exist only as a subjective feel, differing from one person to the

next, about the credibilities of the various hypotheses.

3. Whenever possible, the basic statistical report should be in the form of a *confidence interval*. Briefly, a confidence interval is a subset of the alternative hypotheses computed from the experimental data in such a way that for a selected confidence level α, the probability that the true hypothesis is included in a set so obtained is α. Typically, an α-level confidence interval consists of those hypotheses under which the p value for the experimental outcome is larger than $1-\alpha$ (a feature of confidence intervals which is sometimes confused with their definition), in which case the confidence-interval report is similar to a simultaneous null-hypothesis significance test of each hypothesis in the total set of alternatives. Confidence intervals are the closest we can at present come to quantitative assessment of hypothesis-probabilities (see *technical note*, below), and are currently our most effective way to eliminate hypotheses from practical consideration—if we choose to act as though none of the hypotheses not included in a 95% confidence interval are correct, we stand only a 5% chance of error. (Note, moreover, that this probability of error pertains to the incorrect simultaneous "rejection" of a major part of the total set of alternative hypotheses, not just to the incorrect rejection of one as in the NHD method, and is a *total* likelihood of error, not just of Type I error.) The confidence interval is also a simple and effective way to convey that all-important statistical datum, the conditional probability (or probability density) function—i.e., the probability (probability density) of the observed outcome under each alternative hypothesis—since for a given kind of observed statistic and method of confidence-interval determination, there will be a fixed relation between the parameters of the confidence interval and those of the conditional probability (probability density) function, with the endpoints of the confidence interval typically marking the points at which the conditional probability (probability density) function sinks below a certain small value related to the parameter α. The confidence-interval report is not biased toward some favored hypothesis, as is the null-hypothesis significance test, but makes an impartial simultaneous evaluation of all the alternatives under consideration. Nor does the confidence interval involve an arbitrary decision as does the NHD test. Although one person may prefer to report, say, 95% confidence intervals while another favors 99% confidence intervals, there is no conflict here, for these are simply two ways to convey the same information. An experimental report can, with complete consistency and some benefit, simultaneously present several confidence intervals for the parameter being estimated. On the other hand, different choices of significance level in the NHD method is a clash of incompatible decisions, as attested by the fact that an NHD analysis which simultaneously presented two different significance levels would yield a logically inconsistent conclusion when the observed statistic has a value in the acceptance region of one significance level and in the rejection region of the other.

Technical note: One of the more important problems now confronting theoretical statistics is exploration and clarification of the relationships among inverse probabilities derived from confidence-interval theory, fiducial-probability theory (a special case of the former in which the estimator is a sufficient

statistic), and classical (i.e., Bayes') inverse-probability theory. While the interpretation of confidence intervals is tricky, it would be a mistake to conclude, as the cautionary remarks usually accompanying discussions of confidence intervals sometimes seem to imply, that the confidence-level α of a given confidence interval I should not really be construed as a probability that the true hypothesis, H, belongs to the set I. Nonetheless, if I is an α-level confidence interval, the probability that H belongs to I as computed by Bayes' theorem given an a priori probability distribution will, in general, *not* be equal to α, nor is the difference necessarily a small one—it is easy to construct examples where the a posteriori probability that H belongs to I is either 0 or 1. Obviously, when different techniques for computing the probability that H belongs to I yield such different answers, a reconciliation is demanded. In this instance, however, the apparent disagreement is largely if not entirely spurious, resulting from differences in the evidence relative to which the probability that H belongs to I is computed. And if this is, in fact, the correct explanation, then fiducial probability furnishes a partial solution to an outstanding difficulty in the Bayes' approach. A major weakness of the latter has always been the problem of what to assume for the a priori distribution when no pre-experimental information is available other than that supporting the background assumptions which delimit the set of hypotheses under consideration. The traditional assumption (made hesitantly by Bayes, less hesitantly by his successors) has been the "principle of insufficient reason," namely, that given no knowledge at all, all alternatives are equally likely. But not only is it difficult to give a convincing argument for this assumption, it does not even yield a unique a priori probability distribution over a continuum of alternative hypotheses, since there are many ways to express such a continuous set, and what is an equilikelihood a priori distribution under one of these does not necessarily transform into the same under another. Now, a fiducial probability distribution determined over a set of alternative hypotheses by an experimental observation is a measure of the likelihoods of these hypotheses relative to all the information contained in the experimental data, but based on no pre-experimental information beyond the background assumptions restricting the possibilities to this particular set of hypotheses. Therefore, it seems reasonable to postulate that the no-knowledge a priori distribution in classical inverse probability theory should be that distribution which, when experimental data capable of yielding a fiducial argument are now given, results in an a posteriori distribution identical with the corresponding fiducial distribution.

4. While a confidence-interval analysis treats all the alternative hypotheses with glacial impartiality, it nonetheless frequently occurs that our interest is focused on a certain selection from the set of possibilities. In such case, the statistical analysis should also report, when computable, the precise p value of the experimental outcome, or better, though less familiarly, the probability density at that outcome, under each of the major hypotheses; for these figures will permit an immediate judgement as to which of the hypotheses is most favored by the data. In fact, an even more interesting assessment of the postexperimental credibilities of the hypotheses is then possible through use of "likelihood ratios" if one is willing to put his pre-experimental feelings about their relative likelihoods into a quantitative estimate. For let $Pr(H,d)$, $Pr(d,H)$, and $Pr(H)$ be, respectively, the probability of a hypothesis H in light of the experimental data d (added to the information already available), the probability of data d under hypothesis H, and the pre-experimental (i.e., a priori) probability of H. Then for two alternative hypotheses H_0 and H_1, it follows by classical theory that

$$\frac{Pr(H_0, d)}{Pr(H_1, d)} = \frac{Pr(H_0)}{Pr(H_1)} \times \frac{Pr(d, H_0)}{Pr(d, H_1)} \quad [1][2]$$

[2] When the numbers of alternative hypotheses and possible experimental outcomes are transfinite, $Pr(d, H) = Pr(H, d) = Pr(H) = 0$ in most cases. If so, the probability ratios in Formula 1 are replaced with the corresponding probability-density ratios. It should be mentioned that this formula rather idealistically presupposes there to be no doubt about the correctness of the background statistical assumptions.

Therefore, if the experimental report includes the probability (or probability density) of the data under H_0 and H_1, respectively, and its reader can quantify his feelings about the relative pre-experimental merits of H_0 and H_1 (i.e., $Pr(H_0)/Pr(H_1)$), he can then determine the judgment he should make about the relative merits of H_0 and H_1 in light of these new data.

5. Finally, experimental journals should allow the researcher much more latitude in publishing his statistics in whichever form seems most insightful, especially those forms developed by the modern theory of estimates. In particular, the stranglehold that conventional null-hypothesis significance testing has clamped on publication standards must be broken. Currently justifiable inferential algorithm carries us only through computation of conditional probabilities; from there, it is for everyman's clinical judgment and methodological conscience to see him through to a final appraisal. Insistence that published data must have the biases of the NHD method built into the report, thus seducing the unwary reader into a perhaps highly inappropriate interpretation of the data, is a professional disservice of the first magnitude.

Summary

The traditional null-hypothesis significance-test method, more appropriately called "null-hypothesis decision [NHD] procedure," of statistical analysis is here vigorously excoriated for its inappropriateness as a method of *inference*. While a number of serious objections to the method are raised, its most basic error lies in mistaking the aim of a scientific investigation to be a *decision*, rather than a *cognitive* evaluation of propositions. It is further argued that the proper application of statistics to scientific inference is irrevocably committed to extensive consideration of inverse probabilities, and to further this end, certain suggestions are offered, both for the development of statistical theory and for more illuminating application of statistical analysis to empirical data.

REFERENCE

Braithwaite, R. B. *Scientific explanation.* Cambridge, England: Cambridge Univer. Press, 1953.

(Received June 30, 1959)

Psychological Bulletin
1960, Vol. 57, No. 5, 429–434

GLUTAMIC ACID AND HUMAN INTELLIGENCE[1]

ALEXANDER W. ASTIN[2] AND SHERMAN ROSS
University of Maryland

Glutamic acid administered in supranormal quantities has been reported to enhance the intellectual functioning of mentally defective patients. Interest in this relationship was stimulated in part by the work of Weil-Malherbe (1936), who reported that l(+) glutamic acid was the only one of 12 amino acids studied which was capable of maintaining oxygen uptake in sliced brain tissue. Since that time a major literature has accumulated relating glutamic acid to human intelligence, brain function, epilepsy, audiogenic seizures, and performance of rodents in mazes. In this paper we will review the literature which evaluates glutamic acid therapy as a means of improving intellectual functioning in mental defectives.

Studies of this problem ordinarily employ institutionalized mental defectives who are placed on a high diet of glutamic acid for a specified period of time (usually two to six months). Psychological tests and/or clinical observations of intellectual functioning are made before and after treatment.

Inspection of this literature reveals a set of highly conflicting findings. The picture is somewhat clarified, however, when the studies are classified in terms of (*a*) whether or not positive results were reported, and (*b*) whether or not a control group was used. Table 1 summarizes 33 studies with mentally deficient *S*s in terms of these two variables. A study was classified as positive if significant gains in IQ scores or improvements in "intellectual functioning" were attributed to treatment with glutamic acid. Studies employing "controls" were liberally designated as those in which a similarly diagnosed group was studied in the absence of glutamic acid medication. A chi square with correction for continuity is significant at the .001 level ($\chi^2 = 12.99$), indicating that positive results tend to be related to a lack of controls. Obviously, the crucial studies occur in the Control-Positive cell of Table 1. An attempt will be made here to evaluate these studies according to the following additional methodological considerations: (*a*) adequacy of control group; (*b*) use of placebo; (*c*) control of "environmental stimulation," such that all *S*s are treated similarly except for drug administration; (*d*) ignorance of *S*s and *E*s of the medications or placebos; (*e*) statistical treatment of data; and (*f*) control of taste differences between drug and placebo (some forms of glutamic acid have a markedly bitter taste).

Albert, Hoch, and Waelsch (1946) reported positive clinical results with eight mental defectives serving as their own controls. Intelligence test performance in general seemed to improve during glutamic acid medication and drop during placebo administration. However, some tests showed both gains and losses during acid and placebo periods, and no statistical analysis was attempted.

[1] Now with the Veterans Administration, Baltimore, Maryland.

[2] Supported in part by grant MY-1604 from the National Institute of Mental Health to the University of Maryland.

TABLE 1

33 STUDIES CLASSIFIED IN TERMS OF RESULTS AND USE OF CONTROLS

	Positive Results	Negative Results
Control	Albert, Hoch, & Waelsch (1946, 1951) Foale (1952) Head (1955) Kurland & Gelash (1953) Zimmerman, Burgmeister, & Putnam (1948) $f=6$	Berguis (1954) Ellson, Fuller, & Urmston (1950) Ernsting (1949) Kantor & Boyes (1951) Kerr & Szurek (1950) Loeb & Tuddenham (1950) Lombard, Gilbert, & Donofrio (1955) McCulloch (1950) Milliken & Standen (1951) Oldfelt (1952) Quinn & Durling (1950a) Zabrenko & Chambers (1952) Zublin & Lutz (1953) $f=13$
No Control	de la Fuente Muniz, Zuniga, & Yanowsky (1950) Delay, Pichot, Puech, & Perse (1951) Harney (1950) Hoven (1951) Kane (1953) Levine (1949) Müller (1953, 1954) Quinn & Durling (1950b) Schwobel (1950, 1952) Zimmerman & Burgmeister (1950) Zimmerman, Burgmeister, & Putnam (1949a, 1949b) $f=14$	$f=0$

Moreover, it was not reported if the testors had knowledge of the periods of medication, and there was no control of taste differences between acid and placebo.

The same investigators (1951) reported a more adequately designed follow-up of their first report. Controls were matched on MA, IQ, age, and sex. Significant gains in IQ scores obtained during glutamic acid medication dropped significantly during placebo periods. Testors were not informed of the medication schedules, but " . . . testing of the patients was more difficult during glutamic acid periods as later judged by the number of times it was recorded that a patient had been negativistic, distractible, etc." (p. 487). Obviously the testors possessed information which may have enabled them to differentiate the control and experimental *Ss*. With regard to the taste variable, " . . . rare questions regarding differences in taste were explained as due to differences in the strength of the compound administered" (p. 475). The potential effects of such information upon the parents (who administered the medication) and the children provides an unknown, hence uninterpretable, source of variance.

Zimmerman, Burgmeister, and Putnam (1948), using 69 mentally defective children, reported significant gains in IQ scores during glutamic acid medication. Thirty-seven of these *Ss* served as "controls" in the

sense that they had been given an intelligence test prior to the pretest for the study. The interval between these two testings, however, ranged from six months to eight *years*. Recognizing the potential inadequacy of such controls, the authors wrote, " . . . abrupt environmental changes might possibly influence the intelligence test score. This was not an important factor . . . since . . . the daily pattern of their lives was not appreciably changed" (p. 597). In another paragraph they stated, "Overdoses may produce distractibility . . . insomnia . . . [and] occasional gastric distress" (p. 594). Describing dosage procedures in a preliminary report (1947), "Glutamic acid was administered . . . to the point where increased motor activity was present, or where parents complained about the distractibility and noncooperativeness of the child" (p. 175). It certainly appears as though "abrupt environmental changes" *did* occur at least for some *Ss*, although no method for evaluating their specific contribution in this situation was provided.

A study by Foale (1952) employed two groups of 15 mentally defective boys equated on age, IQ, and length of institutionalization. Gains in IQ scores among the treated *Ss* were attributed to glutamic acid medication, although no placebos were employed for control *Ss*, no control on testor knowledge was reported, and no statistical tests were attempted.

Kurland and Gelash (1953) matched 13 adult male mental defectives with 13 controls on age and IQ scores. Significantly greater gains in IQ scores were reported for the experimental *Ss*. No attempts to control taste differences were made and it was not revealed if the testors knew the medication schedules.

The final and perhaps most carefully designed Control-Positive study was carried out by Head (1955). Three groups of 30 children each (schizophrenics, mental defectives, and normals) were employed in a cross-over procedure which lasted for three, one-month periods. Subgroups of 10 *Ss* received different combinations and orders of medication and placebo. The author's conclusion of a significantly beneficial effect attributable to glutamic acid is based primarily on a comparison of two subgroups of mental defectives before and after their initial month in the study. Mean changes in IQ scores of 6.0 and 3.0 were reported for the experimental and control groups, respectively, during this interval. However, in computing the standard error of this mean difference (p. 120), Head incorrectly employed the standard errors of the two pre- and posttest mean differences, instead of the standard deviations of the two distributions of difference scores. When appropriate substitutions are made, the t ratio ($df = 18$) drops from 3.52 ($p < .01$) to 1.12 ($p > .05$). Head controlled taste differences, but no control over testor knowledge was reported.

The methodological criticisms of these studies do not, of course, render the authors' interpretations invalid; but the amount of contradictory experimental evidence (Control-Negative cell of Table 1) raises considerable doubt. Seven studies (Ellson, Fuller, & Urmston, 1950; Kantor & Boyes, 1951; Kerr & Szurek, 1950; Loeb & Tuddenham, 1950; Lombard, Gilbert, & Donofrio, 1955; Milliken & Standen, 1951; Zabrenko & Chambers, 1952) controlled both testor knowledge and taste differences between placebo and glutamic acid. Six studies (Ellson, Fuller, & Urmston, 1950; Loeb & Tuddenham, 1950; Lombard, et al., 1955; Milliken

& Standen, 1951; Oldfelt, 1952; Quinn & Durling, 1950a) employed matching procedures in the selection of *Ss*.

Some specific points of comparison can be made between certain Control-Positive and Control-Negative studies. The negative study by Kantor and Boyes (1951) was intended as replication of the earlier Albert-Hoch-Waelsch study (1946). Improvements in the design included larger *Ns*, more homogeneous and more reliable diagnoses, control of testor knowledge, and control of taste differences. Two negative studies (Ellson, et al., 1950; Loeb & Tuddenham, 1950) attempted to replicate the positive study by Zimmerman, Burgmeister, and Putnam (1947, 1948). The main improvement in the design of these replications was the use of a control group which was studied concurrently with the experimental group. Ellson, Fuller, and Urmston reported gains in IQ scores for *all Ss* which were comparable to those reported by Zimmerman, Burgmeister, and Putnam during the medication period. Loeb and Tuddenham, in attempting to explain the lack of change in IQ scores found during the six-month to eight-year control period of the Zimmerman-Burgmeister-Putnam study, point out that facilitative practice effects on IQ tests are less likely to appear over such relatively long periods of time.

Three other Control-Negative studies (Ernsting, 1949; Quinn & Durling, 1950a; Zabrenko & Chambers, 1952) also reported significant gains in IQ scores for experimental and control groups. The Zabrenko-Chambers paper studied the "environmental stimulation" variable directly. Transferring mental defectives to a setting characterized by "increased attention and emotional support" was followed by significant increases in IQ scores.

Himwich (1954) has discussed one aspect of this research which has been given little attention, that is, the form in which glutamic acid is usually administered. She points out that some studies have employed the unneutralized acid, while others have used the hydrochloride or salt (usually sodium glutamate). Presenting data from one *S* to demonstrate that the unneutralized acid does not enter the bloodstream as efficiently as the hydrochloride or salt, Himwich concludes, "It is interesting that the majority of favorable reports . . . have been obtained with sodium glutamate or the hydrochloride" (p. 352). Even though some of the papers reviewed are rather obscure on this matter, a check on those listed in Table 1 does not support Himwich's observation. As far as could be determined, all of the studies in the Control-Positive cell of Table 1 used the unneutralized acid. On the other hand, at least three of the Control-Negative papers (Ellson, et al., 1950; Loeb & Tuddenham, 1950; Milliken & Standen, 1951) used sodium glutamate.

Rogers and Pelton (1957) recently reported positive effects on IQ scores using a closely related compound, glutamine. The authors used small doses for a relatively short time (six weeks), and concluded that "some types of mental deficiency are associated with an impaired ability to synthesize glutamine from glutamic acid" (p. 88). Although *Ss* were carefully matched and testor knowledge was controlled, the results were equivocal ($p < .10 > .05$) and no control of taste differences was attempted. These workers plan to publish studies with larger *Ns* in the near future.

Another aspect which has received no systematic study is that of side effects. Very large doses of glutamic

acid can result in flushing (Fincle & Reyna, 1958), nausea and vomiting (Himwich, Wolff, Hunsuker, & Himwich, 1955), in addition to the distractibility, hyperactivity, etc., mentioned earlier. These effects appear to be temporary, and disappear with the cessation or reduction of medication. The importance of such effects for the present review is that their appearance enables observers to differentiate between control and experimental *Ss*. Moreover, the fact that experimental *Ss* who develop such side reactions will require additional care and attention introduces another "environmental stimulation" which may contaminate the findings.

A final point is that of diet. It is likely that the diets of chronic patients in certain institutions are in many respects deficient. The complexity of amino acid metabolism and the difficulty in defining which amino acids are essential for adequate nutrition are well known (Woods, 1950). The effects of glutamic acid medication may depend to a large extent upon the recent dietary history of the population under study.

In conclusion, it appears that a specific effect of supranormal amounts of glutamic acid upon human intelligence has yet to be convincingly demonstrated. The more carefully designed studies tend to be negative almost without exception. Needless to say, the classification of studies in Table 1 provides testimony in support of the demand for controlled experimentation.

Summary

A review of the literature relating glutamic acid medication to the intellectual functioning of mental defectives indicates that positive effects tend to be reported in studies not employing a control group. The few positive studies employing controls contain methodological flaws, rendering their conclusions difficult to accept. The tendency for negative findings to occur in the more adequately designed experiments sheds doubt on the hypothesis that glutamic acid medication has a specifically beneficial effect on intellectual functioning.

REFERENCES

Albert, K., Hoch, P., & Waelsch, H. Preliminary report on the effect of glutamic acid administration in mentally retarded subjects. *J. nerv. ment. Dis.*, 1946, **104**, 263–274.

Albert, K., Hoch, P., & Waelsch, H. Glutamic acid and mental deficiency. *J. nerv. ment. Dis.*, 1951, **114**, 471–491.

Berguis, R. Psychologische Untersuchungen über Wirkungen der glutaminsaüre. *Jb. Psychol. Psychother.*, 1954, **2**, 21–70.

de la Fuente Muniz, R., Zuniga, C., & Yanowsky, Luisa. Accion del acido glutamico sobre el funcionamiento intelectual de los ninos debiles mentalis: Revision de la literatura y estudie de 36 casos. *Rev. Mex. Psiquiat. Neurol. Neurocirug.*, 1950, **1**, 55–62.

Delay, J., Pichot, P., Puech, J., & Perse, J. L'acid glutamique en psychiatrie. *Sem. Hop. Paris*, 1951, **27**, 2143–2154.

Ellson, D. G., Fuller, P. R., & Urmston, R. The influence of glutamic acid on test performance. *Science*, 1950, **112**, 248–250.

Ernsting, W. De behandeling van geestelijk achtergeblevan kindern met glutaminezuur. *Ned. Tijdschr. Geneesk.*, 1949, **93**, 1044–1054.

Fincle, L. P., & Reyna, L. J. A one year study of L-Glutavite on long term hospitalized, elderly, schizophrenic patients. *J. clin. exp. Psychopath.*, 1958, **19**, 7–18.

Foale, Martha. The treatment of mental defectives with glutamic acid. *J. ment. Sci.*, 1952, **98**, 483–487.

Harney, Maureen. *Some psychological characteristics of retarded girls before and following treatment with glutamic acid.* Washington, D. C.: Catholic Univer. Amer. Press, 1950.

Head, M. L. Some metabolic correlates of amentia and dementia: A comparison of the physical, metabolic, and psychological measures of childhood schizophrenics, men-

tal defectives, and normal subjects, and the effect of glutamic acid therapy upon these relationships. Unpublished doctoral dissertation, New York Univer., 1955.

HIMWICH, H. E., WOLFF, K., HUNSICKER, A. L., & HIMWICH, WILLIAMINA A. Some behavioral effects associated with feeding sodium glutamate to patients with psychiatric disorders. *J. nerv. ment. Dis.*, 1955, **121**, 40–49.

HIMWICH, WILLIAMINA A. Absorption of l-glutamic acid. *Science*, 1954, **120**, 351–352.

HOVEN, H. L'acide glutamique chez les deficients mentaux. *Acta neurol. psychiat. Belg.*, 1951, **51**, 1–8.

KANE, ELIZABETH D. Differential indications for the use of glutamic acid. *Amer. J. Psychiat.*, 1953, **109**, 699–700.

KANTOR, R. E., & BOYES, F. E. Effects of supranormal diet of glutamic acid on the test performance of paretics. *Science*, 1951, **113**, 681–682.

KERR, JR., W. J., & SZUREK, S. A. Effects of glutamic acid on mental function. *Pediatrics*, 1950, **6**, 645–647.

KURLAND, A. A., & GELASH, C. A. A study of the effect of glutamic acid on delinquent adult male mental defectives. *Amer. J. ment. Defic.*, 1953, **57**, 669–680.

LEVINE, EDNA S. Can we speed up the slow child? *Volta Rev.*, 1949, **51**, 269–270.

LOEB, H. G., & TUDDENHAM, R. D. Does glutamic acid administration influence mental function? *Pediatrics*, 1950, **6**, 72–77.

LOMBARD, J. P., GILBERT, J. G., & DONOFRIO, A. F. The effects of glutamic acid upon the intelligence, social maturity and adjustment of a group of mentally retarded children. *Amer. J. ment. Defic.*, 1955, **60**, 122–132.

MCCULLOCH, T. L. The effect of glutamic acid feeding on cognitive abilities of institutionalized mental defectives. *Amer. J. ment. Defic.*, 1950, **55**, 117–122.

MILLIKEN, J. R., & STANDEN, J. L. An investigation into the effects of glutamic acid on human intelligence. *J. Neurol. Neurosurg. Psychiat.*, 1951, **14**, 47–54.

MÜLLER, R. Die Wirkung der Glutaminsaüre im Strafvollzug: Ein Beitrag zur Problematik des Erlebniswandels durch Pharmaka. *Psychol. Rdsch.*, 1953, **4**, 200–206.

MÜLLER, R. Über den Erlebniswandel durch Pharmaka. *Z. Psychother. med. Psychol.*, 1954, **4**, 21–33.

OLDFELT, VERA. Experimental glutamic acid treatment in mentally retarded children. *J. Pediat.*, 1952, **40**, 316–323.

QUINN, K. V., & DURLING, DOROTHY. New experiment in glutamic acid therapy: 24 cases classified as mental deficiency, undifferentiated, treated with glutamic acid for six months. *Amer. J. ment. Defic.*, 1950, **55**, 227–234. (a)

QUINN, K. V., & DURLING, DOROTHY. Twelve months' study of glutamic acid therapy in different clinical types in an institution for the mentally deficient. *Amer. J. ment. Defic.*, 1950, **54**, 321–322. (b)

ROGERS, LORENE L., & PELTON, R. B. Effects of glutamine on I.Q. scores of mentally deficient children. *Tex. Rep. Biol. Med.*, 1957, **15**, 84–90.

SCHWOBEL, G. Untersuchungen über die Beeinflussbarkeit psychischer Funktionen durch Glutaminsaüre. *Nervenarzt*, 1950, **21**, 385–393.

SCHWOBEL, G. Untersuchungen über die Wirkung der l+ glutaminsaure auf intellektuelle Funktionen bei debilen und imbezillen Kindern. *Z. Kinderpsychiat.*, 1952, **19**, 101–126.

WEIL-MALHERBE, H. Studies on brain metabolism: Mechanism of glutamic acid in brain. *Biochem. J.*, 1936, **30**, 665.

WOODS, RUTH. Dynamic interrelationships in amino acid nutrition. *Borden's Rev. nutr. Res.*, 1950, **11**, 91–101.

ZABRENKO, R. N., & CHAMBERS, GUINEVERE S. An evaluation of glutamic acid in mental deficiency. *Amer. J. Psychiat.*, 1952, **108**, 881–887.

ZIMMERMAN, F. T., & BURGMEISTER, BESSIE B. The effect of glutamic acid on borderline and high-grade defective intelligence. *NY State J. Med.*, 1950, **50**, 693–697.

ZIMMERMAN, F. T., BURGMEISTER, BESSIE B., & PUTNAM, T. J. A group study of the effect of glutamic acid upon mental functioning in children and adolescents. *Psychosom. Med.*, 1947, **9**, 175–183.

ZIMMERMAN, F. T., BURGMEISTER, BESSIE B., & PUTNAM, T. J. The ceiling effect of glutamic acid upon intelligence in children and in adolescence. *Amer. J. Psychiat.*, 1948, **104**, 593–599.

ZIMMERMAN, F. T., BURGMEISTER, BESSIE B., PUTNAM, T. J. Effects of glutamic acid on the intelligence of patients with mongolism. *Arch. Neurol. Psychiat., Chicago*, 1949, **61**, 275–287. (a)

ZIMMERMAN, F. T., BURGMEISTER, BESSIE B., & PUTNAM, T. J. The effect of glutamic acid upon the mental and physical growth of mongols. *Amer. J. Psychiat.*, 1949, **105**, 661–668. (b)

ZUBLIN, W., & LUTZ, J. Über einen Versuch mit Glutaminsaüre behandelung bei 16, resp. 30 schwachbegabten Schulern. *Z. Kinderpsychiat.*, 1953, **20**, 38–44.

(Received July 7, 1959)

Psychological Bulletin
1960, Vol. 57, No. 5, 435–448

MEASUREMENT OF PERSONALITY AND BEHAVIOR CHANGES FOLLOWING PSYCHOTHERAPY[1]

MELVIN ZAX
University of Rochester
AND ARMIN KLEIN
Residential Treatment Center and University of Rochester

In his general review of the area of psychotherapy in 1946, Snyder (1947) expressed optimism and foresaw that this field was at least in the early stages of becoming a science. He saw as a "commendable trend" the fact that the scientific approach was being more widely used in the study of all methods of therapy and pointed out that the measurement of outcome was undergoing objectification. Since the time of that paper at least 400 studies have been published in which some effort was made to evaluate the effects of psychotherapy. Despite this extensive research activity, there are some (Eysenck, 1952) who have questioned whether anyone has adequately demonstrated that psychotherapy is effective.

Evaluatory research in psychotherapy is a most complex activity but an extremely important one if we are to understand more about the nature of what can bring about personality change. The practical and theoretical problems involved in acquiring *S*s, developing meaningful controls, and making measurements are enormous. Add to these the question of what one should measure, that is, what criterion should be used, and the complexity is increased many times over.

It is the purpose of this paper to summarize and evaluate some of the approaches which have been used to deal with the problem of the criterion. This is based largely on an exhaustive survey of the many experiments, proposals for experiments, theoretical papers, and some reports of case studies involving the evaluation of individual or group psychotherapy which have appeared in the major American psychological and psychiatric journals between 1946 and 1959 (Zax & Klein, 1958). In this context Snyder's (1947) definition of psychotherapy has been adopted which rules out studies devoted to educational procedures and guidance activities emphasizing the giving of information, as well as social activities, occupational therapy, shock therapy, chemotherapy, etc.

The present review is divided into two major sections devoted to (*a*) criteria based on client behavior in the therapy situation or his personal report and (*b*) criteria based on the client's behavior outside of the therapy situation. The studies cited in this paper are selected as being illustrative of these two approaches. Phenomenological measures and indices of client behavior within the therapy situation have been used in some of the major systematic programs for evaluating psychotherapy (Rogers & Dymond, 1954; Snyder, 1953). Measures of extratherapeutic behavior represent logically a most important yardstick. In addition to these two major approaches, psychological tests have also found frequent use as criteria, but, because of

[1] The authors are grateful to E. L. Cowen of the University of Rochester for reading the manuscript and offering many pertinent criticisms and suggestions.

space limitations, it was felt that they might better be reviewed separately.

Intratherapeutic Behavior and Phenomenological Criteria

Criteria based on *S*'s self-experience and his behavior within the therapy situation have stemmed largely from the work of the client centered group who have actively studied their treatment approach. In their research program, they have developed a few instruments which were directly intended to serve as outcome criteria and several indices which have important implications for outcome.

Seeman (1954) constructed a measure which has found considerable use both as a criterion of therapeutic outcome and as a validating instrument for other indices (Rogers & Dymond, 1954). It consists of 10 nine-point scales, several of which required the counselor to evaluate some aspect of the client's experience. This instrument was applied to 23 therapy cases, and it was found that for all items but one there was significant change in the direction of improvement from the beginning to the end of treatment. Also, correlations between ratings on individual items and an item which referred simply to success of outcome revealed that when a client was judged to be successfully treated he was rated highly on scales measuring the extent to which he used therapy as an emotional experience, used it for personal rather than situational exploration, liked and respected his therapist, moved in the direction of both personal integration and situational adjustment, and was satisfied with the outcome of therapy. While this appears as a validation of the instrument it seems likely that the changes measured by many of these individual scales are implicit dimensions of the global judgment of success to which it was compared.

The other instrument which has been used as a criterion in a number of studies (Snyder, 1953) was developed by Tucker (1953) who termed it the "multiple criterion." This involved a Client Post Therapy Scale which was essentially a self-assessment device in which the client was asked to rate his feelings toward such things as the possibility of having problems in the future, the status of the problem which brought him to treatment, relationship with immediate family, sexual adjustment, relationship to others, etc. Another measure as part of the criterion was the Counselor Post Therapy Check List which involved 29 items referring to the client's behavior during therapy and was based on a careful review of therapy notes and interview recordings. This check list was filled out by both the therapist and in each case by one other of a group of trained raters who also used transcribed interview material. Finally, the first and last interviews in each case were analyzed as to the number of positive and negative emotional statements made by the client and an index derived by dividing number of negative statements by the sum of negative and positive ones.

The client's self-report which was an integral part of Tucker's "multiple criterion" has been used as the sole criterion at times and represents the most direct phenomenological measure of therapy outcome. Investigations using such measures have ranged from those employing elaborate rating devices, with some effort at standardizing the procedure, to ratings based on relatively unstandardized interviews in which the *S* is asked to describe his present state or changes which may have oc-

curred as the result of therapy. Fiedler's study (1949) serves as an example of the former. He had *Ss* fill out a 10-item self-rating scale with each item scaled from 0 to 12. The items referred to emotional tensions related to the stress of taking academic examinations and to changes as the result of psychotherapy.

In studies using less systematic self-evaluative techniques, like that of Lipkin (1948), general questions have been asked such as, "What seemed to go on during your visits here?" "How do things look to you now?" Responses were evaluated subjectively and the clients' descriptions of their experience in therapy and its effect on them were seen to confirm the expectations of Rogerian theory.

Cowen and Combs (1950) used a third approach for eliciting the clients' evaluation of therapeutic progress. They conducted open-ended follow-up interviews which were recorded and evaluated by three judges as being "successful, progress, or failure" cases.

Other instruments have been developed which elicit self-descriptions from the client. While such descriptions have not been a direct evaluation of the therapy experience itself, they have implications for the effects of therapy and have been used as outcome measures. In one study, Butler and Haigh (1954) used a *Q* sort involving 100 self-referent statements which had been randomly selected from available therapy protocols. *Ss* were required to sort these to describe themselves as they were at the time on a "like-me" to "unlike-me" continuum. They were further asked to make sortings which would describe their own ideal on a "like-ideal" to "unlike-ideal" continuum. The investigators reported significant increases in the correlation between self and ideal sorts of clients who underwent therapy despite the fact that the same clients failed to show such changes on the same sorts made before and after a waiting period prior to the beginning of therapy. A no-therapy control group also failed to demonstrate such changes. Cartwright (1957) found a significant relationship between success in treatment as rated by the therapist and an increased consistency in the sorting of the Butler-Haigh items when three self-sorts were made each using different people as interacting reference points.

Rosenthal (1955) constructed a Morals Value *Q*-Sort comprising 100 statements which the *S* sorted into two piles as being relatively more or less descriptive of himself. This was administered to the patients in his sample before and after treatment and the therapists involved also made the sort. His findings were that patients judged as improved tended to revise more values in the direction of those of the therapist than the unimproved.

Dymond (1953) selected 74 of the Butler and Haigh items which two non-client-centered psychologists had sorted into two equal piles as being characteristic of the well adjusted on the one hand and of the poorly adjusted on the other. These in turn were given to four other judges who sorted them independently in a similar fashion and a high degree of agreement was found. *Ss* were then given an adjustment score based on how many of either kind of statement appeared on the "like-me" or "unlike-me" sides of their sortings. She found scores on this *Q*-adjustment scale, as it was termed, to move toward good adjustment following therapy (1953, 1954). Cartwright and Roth (1957) found the correla-

tion of a client's self and ideal sorts to be related to the *Q*-adjustment sort and the client's self-rating on the Willoughby Emotional Maturity scale. Although Dymond (1953) had not found differences in the *Q*-adjustment scores after a two-month interval during which her *Ss* were waiting to enter therapy and ultimately did, Grummon (1954) did find significant changes in this type of score among *Ss* who requested treatment but then decided against it when it was available. In this case a two month interval had also elapsed between tests. Dymond (1955) re-examined the *Q* sorts of Grummon's *Ss* and concluded that

> although positive adjustment changes appear to take place in maladjusted persons in the absence of psychotherapy, these are not identical with the changes which occur in equally maladjusted persons who complete therapy (p. 107).

She denied that any "deep" reorganization takes place and saw the improvement as characterized by "a strengthening of neurotic defenses and a denial of the need for help."

A number of studies of personality change as seen in the therapeutic interaction have implications for criteria, especially insofar as these changes have often been related to direct evaluations of outcome. Snyder (1945), following a pioneer investigation of the therapy process by Porter (1943), made the earliest of such studies. He classified client statements into four major categories along the dimension of content significance: descriptions of problems, simple responses asking for advice or accepting or rejecting clarification of feeling, responses showing insight into remedies for a problem, and responses which were unrelated to the principal problem of the client. A second dimension for clients' statements was identified as expressions of feeling and nine categories were set up to classify them. These described attitudes expressed in clients' statements as being positive, negative, or ambivalent with reference to the self, the counselor, or other persons or situations. As a result of his analysis of nearly 10,000 client responses in the 48 interviews he used, Snyder concluded that there was a marked tendency for the client's feelings to change in affective tone from negative to positive. Further he noted that in his attitude toward the counselor the patient was slightly rejecting at first, and indifferent during most of the treatment; but in the last interview or so, there was a marked increase in positive attitudes. He also interpreted his findings as indicating that "clients approaching the end of treatment show an excellent amount of insight into the nature of their problem."

In another of the early studies of personality change with psychotherapy, Raimy (1948) was concerned with changes in self-concept. He analyzed client responses in a set of 14 cases by classifying statements into six categories. These involved self-references which were positive, negative, ambivalent, and ambiguous; statements which did not involve self-references; and nonrhetorical questions. He found that in cases considered successfully treated on the basis of the judgments of the counselor, the supervisor of most of the cases, and Raimy himself, the client went from a preponderance of negative and ambivalent self-references to a preponderance of positive self-references. This was taken to support the hypothesis that in successful therapy a positive change in self-concept took place.

Several measures of client experi-

ence were developed in a series of studies of the process of psychotherapy in a single sample of 10 cases at the University of Chicago. Changes in the clients' experience reflected by these measures were found by Raskin (1949) to be associated with success in therapy as judged by the counselor. Thus, in the more successful cases clients showed an increase in acceptance of, and respect for, self as measured by a scale developed by Sheerer (1949); an increase in positive and objective attitudes directed toward the self as measured by a scale developed by Stock (1949); a tendency toward more mature behavior as judged from the client's own verbalizations in therapy (Hoffman, 1949); and a decrease in defensiveness as measured by Haigh (1949).

In a later study of his own, Raskin developed a four-step scale, illustrated at each point by three examples of client statements, on the basis of which judges estimated whether the client, in what he said, was being governed largely by the expectations of others or by his own values and standards. Ratings on this "locus of evaluation" scale were found to correlate significantly with therapists' ratings as to the success of treatment and with the five parallel interview measures described in the previous study, but not with rated change on the Rorschach.

In a later study of the changes in personality in successful psychotherapy, seen phenomenologically, Vargas (1954) measured self-awareness in three ways and related increase on his measures to a number of criteria of outcome. He summarized his findings by saying:

> The conclusion which seems to follow from these observations is that the hypothesis—increasing self-awareness during therapy correlated with success in therapy—is confirmed when success is measured by instruments which rate highly those changes and states deducible from client centered theory (p. 165).

It should be noted that nearly all of these studies relating personality change in psychotherapy to judgments of the general outcome of treatment involve a certain circularity. In nearly all cases the judgment as to outcome was made by people holding theoretical viewpoints similar to those of the researchers who developed the scales for measuring change. It is, therefore, likely that the two measures were not completely independent.

A few measures of changes in clients' verbal behavior within the therapy interaction have been developed outside of the client-centered framework. One of these was the Discomfort Relief Quotient (henceforth referred to as DRQ) which was first proposed by Dollard and Mowrer (1953). This measure classifies words, clauses, or sentences as to whether they signify discomfort, relief from discomfort, or a neutrality of emotion. To arrive at the quotient the number of discomfort words, clauses, or sentences are divided by this same number plus the number of relief words, phrases or clauses. Thus, the quotient may vary from zero to one, with scores nearer zero representing a preponderance of expressions of relief and those approaching one indicating considerable expression of discomfort. Dollard and Mowrer made no claim that the DRQ measured "success" in treatment. To do this they felt that it must first be related to a reliable measure of "real life success."

Several attempts have been made to validate the DRQ as a measure of success in therapy. Hunt (1949a, 1949b) applied it in a social casework setting and found that changes in

DRQ failed to correlate significantly with judgments of improvement made by case workers. Other studies (Assum & Levy, 1948; Cofer & Chance, 1950; Murray, Auld, & White, 1954) reported analyses of the published protocols of cases presented by the therapist as successful, with two finding the predicted change and the third finding no relationship.

Kauffman and Raimy (1949) derived a related measure from Raimy's self-concept categories (described above). It consisted of the number of negative self-references plus the number of ambivalent self-references divided by the number of negative self-references plus the number of ambivalent self-references plus the number of positive self-references (more conveniently termed the PNAvQ). Using this quotient, they analyzed 17 verbatim interviews and compared their analysis with an analysis of the same protocols using the DRQ. They concluded that both methods traced changes from maladjustment to adjustment in a similar fashion. They also noted that PNAvQ judgments were obtained in about one-third the time required for DRQ judgments.

Another study of the nature of personality change was recently reported by Berg (1958). He analyzed an eight-interview protocol of a case published by Rogers as successful and proposed that early in treatment, clients are preoccupied with self and move in the direction of a more empathic concern for others. He made a frequency count of "ego" words (I, me, my, myself, mine), "empathic" words (we, our, they, us, you, your), "negative" words and "expletive-bombastic sounds" at various points in treatment. It was found that empathic words did indeed increase while ego, negative, and expletive-bombastic expressions decreased with succeeding interviews.

Most recently, Rogers (1958) has developed and given a preliminary report on a scale of process levels in psychotherapy which bears considerable significance for the measurement of the effects of successful psychotherapy. Again, his goal was a further understanding of the nature of change in personality from a theoretical framework rather than measurement of outcome. He conceived that clients move "not from fixity or homeostasis through change to a new fixity . . . but much the more significant continuum is from fixity to changingness. . . ." He hypothesized that the nature of clients' immediate relationship to their feelings at any point in the therapeutic interaction might indicate their position on a seven-stage continuum.

Behavioral Criteria

In many instances, studies of the results of psychotherapy have used criteria which depend on an evaluation of the way the patient actually behaves without inference as to its personal meaning for him. Such indices were generally developed directly as criteria for use in a given situation and were not related to a theoretical framework about personality change.

Of the many studies which have used behavioral criteria, certain ones have been particularly noteworthy in that they dealt with crucial aspects of behavior which can be objectively established. The simplest of such criteria focused on relatively circumscribed individual behaviors which were seen to be central to the person's difficulty in living. The more complex criteria attempted to assess wide, more representative areas of functioning through the use of elaborate rating scales.

A study by Friedman (1950) is typical of those employing criteria

emphasizing delimited behaviors which are central to the person's difficulty in living. His *Ss* were 50 patients complaining of a "phobia of travel," which can be objectively measured. Evaluation was based on their ability to travel after treatment and it was found that 12 patients were unimproved, 15 showed some improvement, and 23 were completely recovered. Another example of a study utilizing a single symptom which bore implications for a much wider range of behavior was that of Teuber and Powers (1951). They simply totaled the number of court appearances among a large group of potential juvenile delinquents who had received treatment and made comparisons with a matched control group which had received no treatment. No significant differences were found between groups on this measure.

A variation in the use of an important individual behavior as a criterion was introduced by Thetford (1952) who derived an autonomic measure of frustration tolerance. This study stands out in that the behavior which was measured was not a specific complaint but depended on the theoretical consideration that therapy should reduce anxiety and tension so that the manner in which one responds to stress as reflected in the autonomic nervous system should be altered. He developed a "Recovery Quotient" based on various Galvanic Skin Response measures and found significant changes as the result of psychotherapy which indicated the development of a higher frustration threshold.

The criterion used by Pascal and Zax (1956) likewise involved objective behavioral measures, but these varied with the individual patient, reflecting presenting complaints. These complaints were evaluated for 30 cases which had undergone various types of treatment. In 28 of these, changes in the predicted direction were found.

Institutional settings have made it possible to study wider samples of behavior objectively. In such settings Cowden, Zax, Hague and Finney (1956); Fox (1954); and Ludwig and Ranson (1947) have used multiple but individually significant behaviors as their criteria. Cowden et al. (1954) considered the number of times hospitalized patients required neutral wet packs, electroconvulsive maintenance shock, or engaged in fights, in addition to such indications of improvement as transfer to a ward requiring a higher level of integration or discharge from the hospital. They concluded that patients who received group psychotherapy in addition to tranquillizing drugs showed more improvement than various control groups. To evaluate the effects of counselling programs in a prison, Fox (1954) used such behavioral criteria as work stability, school stability, financial budgeting, reports from chaplain, block officers, and work supervisors, successful discharge from parole, and return to prison as a parole violator; he found counseled groups had significantly higher adjustment scores on such indices than similar uncounseled groups. In a report of results of psychiatric treatment among soldiers, Ludwig and Ranson (1947) reported that relatively high percentages of treated patients were able to return to duty stations and that ratings of commanding officers indicated that most of them were able to perform their services adequately.

Many studies have made use of longer and more elaborate rating scales which attempted to assess the extratherapy functioning of the individual on the basis of diverse be-

havioral observations. One of the older instruments of this type which was used in the evaluation of treatment with children (Gersten, 1951; Mehlman, 1953) is the Haggerty-Olson-Wickman Behavior Rating Schedules (Jones, 1941). This consists of two separate schedules (A and B) the first of which (A) lists 15 problems such as cheating, lying, defiance of discipline, speech difficulties, sex offenses, obscene notes, talk, or pictures, etc. Raters checked in one of four columns according to the frequency of occurrence of each for a given individual. Standardized weights were assigned according to the frequency and seriousness of a given problem. The other schedule (B) comprised a series of 35 graphic five-point rating scales covering traits which may be classified according to intellectual, physical, social, and emotional traits. On the basis of ratings made before and after group therapy with juvenile delinquents, Gersten (1951) reported progress in emotional security and social maturity among his subjects. Mehlman (1953), who used the scale to rate mentally retarded children before and after group therapy, found significant increases in adjustment at the time of the second rating.

Of the many devices which have been used to evaluate change in hospitalized patients, perhaps the most promising and certainly the most searching are the Palo Alto Hospital Adjustment Scale (McReynolds & Ferguson, 1953) and the Lorr Multidimensional Scale (Lorr, 1953). The Palo Alto scale consists of 90 descriptive statements applicable to psychiatric patients. Examples of these statements are, "the patient ignores the activities around him" or "the patient's talk is mostly not sensible." Each one is marked as true, not true, or does not apply, for a particular patient and is keyed in such a manner that a general hospital adjustment score can be obtained. The scale was designed to be filled out at intervals by ward personnel who are familiar with the patient's behavior. On this measure, schizophrenics were seen to improve with group psychotherapy (Semon & Goldstein, 1957). In another study (Wilcox & Guthrie, 1957) items from this scale were combined with others suggested by personnel in an institution for defective children, and by this index group therapy was found to be effective.

The Lorr scale consists of 62 brief rating scales which are directed toward observable or inferable patient behavior. Many of the items refer to relatively objective behaviors concerning which judgments should be quite reliable, such as bizarre postures, speech peculiarities, orientation, eating, sleeping, assaultiveness. On the other hand many other items refer to aspects of behavior which are probably less reliably rated such as emotional responsiveness, attitude toward himself, suspiciousness, recurrence of useless thoughts, etc. The use of this scale was reported in a study with long term schizophrenic patients who were seen by this measure to have improved significantly more than a control group (Funk, Shatin, Freed, & Rockmore, 1955).

The scales used to evaluate outpatients have as a rule been more difficult to apply and often have been more complex. This is due to the obvious fact that the behavior of the nonhospitalized patient is less limited by the structured aspects of institutional life so that he functions in a much wider range. Observation is thereby also made more difficult. Hunt (1949a, 1949b) has attempted

to measure "movement" in social casework by developing a criterion to evaluate the DRQ. Movement was defined as "the change which appears in an individual client and/or his environmental situation between the opening and closing of his case" (Hunt, 1949b, p. 76). His scale was set up in seven steps ranging from minus two, through zero, to plus four with anchoring illustrations at each of these three points. It was found that experienced workers could use the scale reliably, but no relationship was found between movement and DRQ changes in the course of therapy.

The Willoughby Emotional Maturity Scale (1931) has been used by Rogers (1954) to evaluate changes in psychotherapy. It consists of 60 statements descriptive of varying levels of maturity of functioning. The levels had been determined by 100 clinicians who sorted a large number of statements along a nine-step continuum. The 60 items selected for the scale were representative of the nine levels of maturity and were ones on which there was high agreement among judges. In Rogers' study, each client was rated by himself and two personal friends whom he designated. Although intrarater reliability was high, interrater reliabilities were all low. Conceivably, this scale might have higher reliability in the hands of trained observers although this might limit its use to a somewhat standardized setting such as a dormitory or school setting.

Miles, Barrabee, and Finesinger (1951) developed a series of five-point scales covering the general areas of (*a*) symptoms; (*b*) social adjustment including functioning in the areas of occupation, marriage, interpersonal relations, and sex; (*c*) insight; and (*d*) life situation since hospitalization. As a group, the scales were comprehensive and individual steps were well described. On the basis of these instruments, overall evaluations were made of patients and summarized in the categories "apparently recovered, much improved, improved, slightly improved, unimproved, and worse." In using this measure to assess a group of 62 cases two years after treatment, they found that 58% had improved in varying degrees while 42% were unchanged. Imber, Frank, Nash, Stone, and Gliedman (1957) derived a Social Ineffectiveness score on the basis of a series of six-point scales which applied to each of 15 behavioral categories concerning the patient's relationships with the significant individuals in his life (spouse, sibs, children, parents, boss, etc.) Some of the categories were overly independent, withdrawn, superficially sociable, extrapunitive, officious, impulsive, etc. Using this scale they investigated the relationship between improvement and amount of therapeutic contact, and they found less improvement for patients with restricted therapy contacts than for those with more frequent ones.

Raush, Dittman, and Taylor (1959) have made a recent contribution to the methodology of making observations and developing behavioral criteria to assess change with treatment. Working in a residential treatment setting for children, they standardized their observations of six male *Ss* and systematically studied samples of their behavior in a variety of settings including mealtimes, play periods, and an arts-and-crafts period. One set of observations was made early in the children's stay at the center and another 18 months later focusing on interpersonal behavior at these two points in time.

Objective observations were recorded and later rated on a scale based on two polar coordinates: love (affiliate, act friendly) to hate (attack, act unfriendly) and dominate (command, high status action) to submit (obey, low status action). More striking changes were found in the relationships of these children to adults than in their relationships to their peers.

Discussion

As is the case with any measure of personality, a criterion for evaluating the effects of psychotherapy must satisfy the requirements of reliability and validity. The latter usually poses the more serious problem in that no absolute state of complete validity exists as a standard. In dealing with this problem we generally conclude that a given measure is valid for certain specified purposes and not necessarily valid for others. Therefore, we may have a variety of "valid" measures of the outcome of psychotherapy. The judgment of whether these are useful measures, however, must be based upon our evaluation of the purposes for which they are valid. The criteria which have been reviewed will be considered in the light of such issues.

Perhaps the simplest and most direct means of assessing a client's progress in treatment is to ask him to evaluate his own status. Such a phenomenological approach has often been used. Unfortunately, on close analysis, this deceptively simple procedure is seen to be fraught with serious pitfalls. Standards for such assessments will vary both among clients and between client and researcher; clients will vary in the extent to which they can report what they feel; the reports of many clients will be subject to various unconscious distortions; finally, the client's evaluation of his condition may be affected by conscious or semiconscious motives. In positing the "hello-goodbye" effect, Hathaway (1948) has warned of the subtle social influences which limit the reliability of many of the phenomenological measures which have been made. On entering treatment the client is under the conventional pressure to justify his appeal for help so that problems are discussed freely. When seeking to terminate, however, he feels an obligation, out of courtesy toward one who has attempted to help, to express gratitude and satisfaction. A fundamental weakness of the phenomenological approach would, therefore, seem to reside in the difficulty in obtaining reliable assessments. It seems likely that the content of such assessments depends greatly upon who asks for it and the circumstances under which it is requested.

Intratherapy behavior, usually verbal behavior, lends itself to measurement and has been used often as a criterion. In many of the studies reporting the use of such criteria a single theoretical system, that of Rogers, has guided the expectations of researchers. As a result many of these studies relate to each other in a more systematic fashion than is usually the case with outcome studies. The aspects of verbal behavior which have been studied by the client-centered group have usually been carefully defined and found to be reliably measured. Designed to explore personality changes during psychotherapy rather than to be evaluators of psychotherapy, their significance for outcome measures is mostly by implication for they remain unvalidated, not yet having been compared to an independent criterion. Used for the purpose of exploring changes, they

were compared in the published studies only to a judgmental criterion of the therapist who shared the same theoretical point of view as the researcher and whose global judgment could have included the concept under study.

Those intratherapy criteria which have not stemmed from the work of the client centered group have found relatively infrequent use and the one attempt to relate change in DRQ to an independent, external criterion (Hunt, 1949a, 1949b) resulted in an insignificant correlation.

The most serious failing at this time in the use of phenomenological measures and measures of intratherapy behavior as criteria of outcome is that neither has yet been related to everyday, externally observable behaviors in the life space of the *S*s. Unless phenomenological changes and changes in verbal behavior in therapy can be related to concomitant behavioral changes in the family and the community their significance remains unclear.

External measures of clients' behavior stand out as potential criteria having validity for purposes which are extremely important. However, when one attempts to use such criteria he is beset by a host of measurement problems which are much more difficult to resolve than is the case with phenomenological and intratherapy indices. The central problem here is the development of criteria of sufficient breadth that they are meaningful and representative of a wide range of functioning and yet, at the same time, circumscribed enough to be measured with reliability.

The present review would seem to indicate that the development of such criteria is in the stage of infancy. Many workers have been able to reliably observe narrow aspects of functioning which had implications for a wider range of behavior. In such cases, however, the possibility remains that one circumscribed symptom was abandoned in favor of another which was equally or even more disabling. The assessment of broader areas of functioning has been carried on primarily within the confines of institutional settings where the patients' range of functioning is limited. Perhaps the most glaring weakness in the way such criteria have been developed and applied is that there has been no unifying set of principles to guide observations. Consequently, the results which have been reported are fragmented. We are told of a variety of behavioral changes which take place as the result of therapy but very few of these appear in any one study and even fewer are observed in more than one study. It would seem that the present need is for the development of a theory or even a set of loose hypothetical notions about "normal" behavior to guide our observations and systematize our thinking.

It seems likely that one of the obstacles to the development of such a theory has been the reluctance of many psychologists to become embroiled in the philosophical issues of the desirability of different behaviors. Actually, the problem of making value judgments when one conducts research cannot be avoided. The very selection of the phenomena which will be observed and measured is in itself a judgment depending upon the values one holds. Indeed then, the further development of criteria for evaluating the effects of psychotherapy awaits the clarification, resolving, and communication of the values we hold.

One approach to the development

of a systematic set of values which may clarify our thinking about what behavior is generally considered "psychologically desirable" would be to formalize the notion of the client's relationship to social norms which was discussed by Pascal and Zax (1956). Their concern was with the behaviors on the part of the person presenting himself for treatment which were notably deviant from expected social norms (i.e. overt homosexual acts, frequent crying spells, few friends) and the extent to which such behaviors were changed. Other writers have suggested that the clinician does generally function with a concern for such social norms. As the result of his work in the area of personality assessment, Edwards (1957) has suggested that the notions of the clinician about what constitutes disturbance in patients may correspond essentially to an operational definition of what is socially undesirable. Cowen (in press) who was investigating the social desirability variable in personality assessment actually provided data which lends support to this idea. He found a correlation of −.917 between the published ratings of a group of clinicians on 77-trait descriptive terms scaled for abnormality and the social desirability ratings of the same terms by 67 undergraduate students of psychology.

This approach suggests that the person considered most psychologically handicapped is the one who is unable to function in the way in which it is expected that he should in his social group. Furthermore, it suggests that phenomenological reports of how one feels are characteristically considered in the light of such evidence about how he functions in the same way that the felt experience of physical comfort or discomfort is evaluated on the basis of various measures of bodily function.

While this approach, which is probably implicit in the thinking and functioning of most clinicians, may provide a useful beginning to the development of criteria of what therapy should accomplish, it is unlikely that any single set of norms would apply to all. In essence, we are proposing that there are, contentwise, many "normal" or "healthy" personalities. That which is common to each is the ability to function in relation to the norms of his particular social setting. The uniqueness of each individual's social setting makes this a complex area of study and is undoubtedly discouraging. It may well develop, however, that what people have in common is important enough to permit the development of a relatively limited number of norms reflecting basic interpersonal environments which can be useful. At any rate, it would seem that what is now needed is a series of broad normative studies of a personal-social psychological nature. In addition to providing norms which can be used as a foundation for behavioral criteria of "normality," they would provide a basis for determining just which dimensions of social group membership have significance for actual functioning. The availability of a criterion based on such indices would also provide a context in which to evaluate the significance of changes in the experiencing of *S*s, either reported directly or reflected in their intratherapy verbal behavior. Ultimately, a combined measure of related changes in observed behavior and experiencing might facilitate a common, communicative frame of reference among workers of different orientations and be a basis for delineation of dimensions of personality change.

REFERENCES

ASSUM, A. L., & LEVY, S. J. Analysis of a nondirective case with followup interview. *J. abnorm soc. Psychol.*, 1948, **43**, 78–89.

BERG, I. A. Word choice in the interview and personal adjustment. *J. counsel. Psychol.*, 1958, **5**, 130–135.

BUTLER, J. M., & HAIGH, G. V. Changes in the relation between self concepts and ideal concepts consequent upon client-centered counseling. In C. R. Rogers & R. F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago, Ill.: Univer. Chicago Press, 1954.

CARTWRIGHT, D. S., & ROTH, I. Success and satisfaction in psychotherapy. *J. clin. Psychol.*, 1957, **13**, 20–26.

CARTWRIGHT, ROSALIND D. Effects of psychotherapy on self consistency. *J. counsel. Psychol.*, 1957, **4**, 15–22.

COFER, C. N., & CHANCE, JUNE. The discomfort-relief quotient in published cases of counseling and psychotherapy. *J. Psychol.*, 1950, **29**, 219–224.

COWDEN, R. C., ZAX, M., HAGUE, J. R., & FINNEY, R. C. Chlorpromazine: Alone and as an adjunct to group psychotherapy in the treatment of psychiatric patients. *Amer. J. Psychiat.*, 1956, **112**, 898–902.

COWEN, E. L. The social desirability of trait descriptive terms: Preliminary norms and sex differences. *J. soc. Psychol.*, in press.

COWEN, E. L., & COMBS, A. W. Follow-up study of 32 cases treated by nondirective psychotherapy. *J. abnorm. soc. Psychol.*, 1950, **45**, 232–258.

DOLLARD, J., & MOWRER, O. H. A method of measuring tension in written documents. In O. H. Mowrer (Ed.), *Psychotherapy theory and research*. New York: Ronald, 1953.

DYMOND, ROSALIND F. An adjustment score for Q sorts. *J. consult. Psychol.*, 1953, **17**, 339–342.

DYMOND, ROSALIND F. Adjustment changes over therapy from self-sorts. In C. R. Rogers & R. F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago, Ill.: Univer. Chicago Press, 1954.

DYMOND, ROSALIND F. Adjustment changes in the absence of psychotherapy. *J. consult. Psychol.*, 1955, **19**, 103–107.

EDWARDS, A. E. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.

EYSENCK, H. J. The effects of psychotherapy: An evaluation. *J. consult. Psychol.*, 1952, **16**, 319–324.

FIEDLER, F. E. An experimental approach to preventive psychotherapy. *J. abnorm. soc. Psychol.*, 1949, **44**, 386–393.

FOX, V. The effects of counseling on adjustment in prison. *Soc. Forces*, 1954, **32**, 285–289.

FRIEDMAN, J. H. Short-term psychotherapy of "phobia of travel." *Amer. J. Psychother.*, 1950, **4**, 259–278.

FUNK, I. C., SHATIN, L., FREED, E. X., & ROCKMORE, L. Somato-psychotherapeutic approach to long-term schizophrenic patients. *J. nerv. ment. Dis.*, 1955, **121**, 423–437.

GERSTEN, C. An experimental evaluation of group therapy with juvenile delinquents. *Int. J. group Psychother.*, 1951, **1**, 311–318.

GRUMMON, D. L. Personality change as a function of time in persons motivated for therapy. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago, Ill.: Univer. Chicago Press, 1954.

HAIGH, G. Defensive behavior in client-centered therapy. *J. consult. Psychol.*, 1949, **13**, 181–189.

HATHAWAY, S. R. Some considerations relative to nondirective psychotherapy as counseling. *J. clin. Psychol.*, 1948, **4**, 226–231.

HOFFMAN, A. E. A study of reported behavioral changes in counseling. *J. consult. Psychol.*, 1949, **13**, 190–195.

HUNT, J. McV. The problem of measuring the results of psychotherapy. *Psychol. Serv. Cent. J.*, 1949, **1**, 122–135. (a)

HUNT, J. McV. A social agency as the setting for research. *J. consult. Psychol.*, 1949, **13**, 68–91. (b)

IMBER, S. D., FRANK, J. D., NASH, E. H., STONE, A. R., & GLIEDMAN, K. H. Improvement and amount of therapeutic contact: An alternative to the use of no-treatment controls in psychotherapy. *J. consult. Psychol.*, 1957, **21**, 309–315.

JONES, H. E. Haggerty-Olson-Wickman Behavior Rating Schedules. In O. K. Buros (Ed.), *The 1940 mental measurements yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. Pp. 1222–1223.

KAUFFMAN, P. E., & RAIMY, V. C. Two methods of assessing therapeutic progress. *J. abnorm. soc. Psychol.*, 1949, **44**, 379–385.

LIPKIN, S. The client evaluates nondirective psychotherapy. *J. consult. Psychol.*, 1948, **12**, 137–146.

LORR, M. Multidimensional scale for rating psychiatric patients. *VA tech. Bull.*, 1953, TB 10-507, 1–44.

LUDWIG, A. O., & RANSON, S. W. A statistical follow-up of effectiveness of treatment of combat-induced psychiatric casualties: I. Returns to full combat duty. *Milit. Surg.*, 1947, **100**, 51–62.

MCREYNOLDS, P., & FERGUSON, J. T. *Clinical manual for the Hospital Adjustment Scale.* Stanford, Calif.: Stanford Univer. Press, 1953.

MEHLMAN, B. Group therapy with mentally retarded children. *J. abnorm. soc. Psychol.*, 1953, **48**, 53–60.

MILES, H. W., BARRABEE, EDNA L., & FINESINGER, J. E. Evaluation of psychotherapy. *Psychosom. Med.*, 1951, **13**, 83–105.

MURRAY, E. J., AULD, F., JR., & WHITE, ALICE M. A psychotherapy case showing progress but no decrease in the discomfort-relief quotient. *J. consult. Psychol.*, 1954, **18**, 349–353.

PASCAL, G. R., & ZAX, M. Psychotherapeutics: Success or failure. *J. consult. Psychol.*, 1956, **20**, 325–331.

PORTER, E. H., JR. The development and evaluation of a measure of counseling interview procedures: II. The evaluation. *Educ. psychol. Measmt.*, 1943, **3**, 215–238.

RAIMY, V. C. Self reference in counseling interviews. *J. consult. Psychol.*, 1948, **12**, 153–163.

RASKIN, N. J. An analysis of six parallel studies of the therapeutic process. *J. consult. Psychol.*, 1949, **13**, 206–221.

RAUSH, H. L., DITTMAN, A. T., & TAYLOR, T. J. The interpersonal behavior of children in residential treatment. *J. abnorm. soc. Psychol.*, 1959, **58**, 9–26.

ROGERS, C. R. Changes in the maturity of behavior as related to therapy. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change.* Chicago, Ill.: Univer. Chicago Press, 1954.

ROGERS, C. R. A process conception of psychotherapy. *Amer. Psychologist*, 1958, **4**, 142–149.

ROGERS, C. R., & DYMOND, ROSALIND F. (Eds.) *Psychotherapy and personality change.* Chicago, Ill.: Univer. Chicago Press, 1954.

ROSENTHAL, D. Changes in some moral values following psychotherapy. *J. consult. Psychol.*, 1955, **19**, 431–436.

SEEMAN, J. A study of the process of nondirective therapy. *J. consult. Psychol.*, 1949, **13**, 157–169.

SEEMAN, J. Counselor judgments of therapeutic process and outcome. In C. R. Rogers & R. F. Dymond (Eds.), *Psychotherapy and personality change.* Chicago, Ill.: Univer. Chicago Press, 1954.

SEMON, R. G., & GOLDSTEIN, N. The effectiveness of group psychotherapy with chronic schizophrenic patients and an evaluation of different therapeutic methods. *J. consult. Psychol.*, 1957, **21**, 317–322.

SHEERER, ELIZABETH T. The relationship between acceptance of self and acceptance of others. *J. consult. Psychol.*, 1949, **13**, 169–175.

SNYDER, W. U. An investigation of the nature of nondirective psychotherapy. *J. gen. Psychol.*, 1945, **33**, 193–223.

SNYDER, W. U. The present status of psychotherapeutic counseling. *Psychol. Bull.*, 1947, **44**, 297–386.

SNYDER, W. U. *Group report of a program of research in psychotherapy.* State College, Pa.: Pennsylvania State Coll., 1953.

STOCK, DOROTHY. The self concept and feeling toward others. *J. consult. Psychol.*, 1949, **13**, 176–180.

TEUBER, H. L., & POWERS, E. Evaluating therapy in a delinquency prevention program. *Res. Publ. Ass. Nerv. Ment. Dis.*, 1951, **31**, 138–147.

THETFORD, W. N. An objective measurement of frustration tolerance in evaluating psychotherapy. In W. Wolff & J. A. Precker (Eds.), *Success in psychotherapy.* New York: Grune & Stratton, 1952.

TUCKER, J. E. Measuring client progress in client-centered therapy. In W. U. Snyder (Chmn.), *Group report of a program of research in psychotherapy.* State College, Pa.: Pennsylvania State Coll., 1953.

VARGAS, M. J. Changes in self-awareness during client-centered therapy. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change.* Chicago, Ill.: Univer. Chicago Press, 1954.

WILCOX, G., & GUTHRIE, G. Changes in adjustment of institutionalized female defectives following group psychotherapy. *J. clin. Psychol.*, 1957, **13**, 9–13.

WILLOUGHBY, R. R. A scale of emotional maturity. *J. soc. Psychol.*, 1931, **3**, 131–136.

ZAX, M., & KLEIN, A. The criterion in evaluation studies of psychotherapy. Unpublished manuscript, Univer. of Rochester, 1958.

(Received August 17, 1959)